PARSING PSYCHOLOGY: STATISTICAL AND COMPUTATIONAL METHODS USING PHYSIOLOGICAL, BEHAVIORAL, SOCIAL, AND COGNITIVE DATA

EDITED BY: Pietro Cipresso and Jason C. Immekus

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

PARSING PSYCHOLOGY: STATISTICAL AND COMPUTATIONAL METHODS USING PHYSIOLOGICAL, BEHAVIORAL, SOCIAL, AND COGNITIVE DATA

Topic Editors:
Pietro Cipresso, Italian Auxological Institute (IRCCS), Italy
Jason C. Immekus, University of Louisville, United States

Table of Contents

![frontiers in Psychology]

Check for updates

Editorial: Parsing Psychology: Statistical and Computational Methods Using Physiological, Behavioral, Social, and Cognitive Data

Jason C. Immekus[1]* and Pietro Cipresso[2,3]

[1] Educational Leadership, Evaluation and Organizational Development, University of Louisville, Louisville, KY, United States,
[2] Department of Psychology, Universitá Cattolica del Sacro Cuore, Milan, Italy, [3] Applied Technology for Neuro-Psychology
Lab, IRCCS Istituto Auxologico Italiano, Milan, Italy

Editorial on the Research Topic

Parsing Psychology: Statistical and Computational Methods Using Physiological, Behavioral, Social, and Cognitive Data

Advancements in statistical and computational methods have much to contribute to promoting new discoveries in psychology. In particular, as found in diverse disciplines as marketing, health care, and medicine, machine learning and big data offer similarly unique potential to inform, test, and advance our knowledge of mental processes, brain functioning, and behavior in ways that were not previously possible. Psychological data science represents the integration of data science into psychology to explore the abundant information provided by physiological, behavioral, social, and cognitive data alone and together. The merging of psychology and data science is due to a host of factors, including, for example, more readily accessible data, cheaper and more powerful computational processing, and larger data storage capacity. As a result, social science researchers are well-positioned in their ability to incorporate data science into their work to advance and promote our understanding of key factors associated with important psychological outcomes (e.g., mental health).

This Research Topic represents a collection of papers that demonstrate the use, or potential, of psychological data science to pursue both hypothesis-driven experiments and bottom-up data exploration through modern statistical and computational models. As these studies illustrate, machine learning algorithms are already expanding our knowledge of key factors associated with designated outcomes (e.g., academic and brain-based disorders) or have the potential to build on existing and emerging areas of research (e.g., scale development). Broadly, the papers comprising this Research Topic fall into three general areas as related to the application of data science in psychological and psychometric research, including: Supervised and Unsupervised Machine Learning Computing, Contemporary Psychometrics, and Emerging Areas of Research.

SUPERVISED AND UNSUPERVISED MACHINE LEARNING COMPUTING

Machine learning represents a branch of artificial intelligence in which statistical model building is based on automatically implemented algorithms to identify data patterns for classification and

prediction purposes. While there are several categories of machine learning, supervised, and unsupervised learning are, at this time, most commonly found in psychological research. Specifically, supervised learning encompasses the development of statistical models in which user-specified inputs (e.g., predictor variables) and outputs (e.g., outcome variables) are provided, and include classification (e.g., decision tree) and regression (e.g., ordinary least squares and elastic net) techniques. On the other hand, unsupervised learning includes algorithms in which only input data (e.g., variables) are analyzed, with the aim of identifying meaningful patterns in the data, such as cluster analysis (e.g., *K*-means) and dimensionality reduction (e.g., principal components analysis) techniques.

The papers within this section demonstrate the use of supervised and unsupervised learning procedures to address a range of substantive topics in psychological research. Specificity, Zhao et al. apply unsupervised clustering methods using static and dynamic resting-state functional magnetic resonance imaging to four different datasets for the classification of different brain-based disorders (e.g., Alzheimer's Disease and Post-Traumatic Stress Disorder) to compare diagnosis based on traditional approaches that rely on clinical interviews and behavioral assessments. In another study, Guixeres et al. compared two types of neural networks, Multi-layer Perceptron and Radial Basis Function, to determine the use of three neurophysiological measures (i.e., brain response, heart rate variability, and eye tracking) to predict the effectiveness of an online advertisement within an online video platform (i.e., YouTube). Situated in a dynamic network framework, Hasmi et al. apply network models using multilevel time-lagged regression to examine if genetic liability to psychopathology and childhood trauma relate to the network structure of various emotions (e.g., cheerful and relaxed), based on real-time data collected on pairs of sibling and twins using the experience sampling method (ESM; i.e., structured diary technique to record emotions). In consideration of the use of big data to understand how individuals use of technology to ways that lead to positive engagement and peace, Guadagno et al. present a theoretical framework for standardizing Peace Data (e.g., group identity) based on hypothetical and real-data. Specifically, real-data was subjected to a social network analysis to examine gender differences (e.g., connections and likes) in social media use within the context of a large financial institution over a 6-month period.

Additional studies applied machine learning within educational settings that have implications to psychological and educational assessment. Based on TIMSS 2011 data, Yoo uses elastic net to identify a prediction model of Korean 4th graders mathematics achievement using 162 teacher and student variables, in which 12 student and 5 teacher variables are identified as significant predictors. The study also examines the process of scale development from a machine learning perspective. Polyak et al.'s innovative research in computational psychometrics demonstrate the use of machine learning to assess middle school students' collaborative problem-solving (CPS) sub-skills (e.g., evaluate

and maintaining a shared understanding), based on their performance within a virtual, online game in which the player must collaborate with a virtual "agent" with the information to "win" the game. Uniquely, this research demonstrates the ways in which traditional test theory practices can be integrated into assessment design and machine learning algorithms to develop psychometric models to establish classification and measurement rules to assess important student traits.

CONTEMPORARY PSYCHOMETRICS

Studies within this section demonstrate the contribution of current and emerging psychometric analyses to the development and validation of instruments to assess a range of individual traits. In general, the psychometric models and analyses used across the majority of studies have a long, rich history in instrument development (e.g., factor analysis). Their application within these studies represent efforts to address key literature gaps in the assessment of clinically important traits, including the utility of virtual reality to promote the ecological validity of neuropsychological testing. Collectively, the papers serve to highlight the contribution of psychological data science to instrument design and validation.

The papers in this section illustrate the use of contemporary psychometrics and technology to advance our assessment of psychological constructs. For instance, Maldonato et al.'s research focuses on the development of a reduced version of the Temperament and Character Inventory for use among clinical samples, and using logistic regression to examine the predictive validity of scores to determine the presence/absence of psychiatric Axis pathologies. In response to lack of available measures, Olivencia-Carrión et al. present their work on the development and validation, via confirmatory factor analysis, of a measure of mobile phone abuse among young Spanish-speakers in Spain. Vallejo et al. use exploratory factor analysis to examine the dimensionality of the Perceived Stress Scale and, subsequently, examine perceived stress throughout three European countries (Great Britain, France, Spain), based on a large dataset ($N = 37,451$) obtained from a smoking cessation program. To promote the ecological assessment of visual attention, Gamito et al.'s pilot study demonstrates the rich potential of virtual reality technology as an alternative to noted limitations of traditionally administered neuropsychological tests. Verhagen et al.'s study focuses on the assessment of momentary reward-related Quality of Life, based on data collected via ESM among individuals with severe mental illness, which they use to generate a momentary rQoL statistic and, second, conduct a Monte Carlo simulation study to determine four outcomes (e.g., required sample size to reliably measure individuals' behavior setting). Last, Provenzi et al.'s paper presents efforts to develop an observational approach to measure emotional stress regulation among preschool aged children in which they identify and describe the theoretical and methodological considerations in the selection of stress-related episodes.

EMERGING AREAS OF RESEARCH

The papers comprised in the previous sections demonstrate the integration of computational and statistical techniques into psychological and psychometric research. Contrary, the papers in this section address substantive topics in psychological research that provide a foundation for future research based on the application of psychological data science. In particular, the common thread across the subsequent studies is the rich potential that technology affords in data collection and analysis, as well as intervention delivery.

The studies represented in this section provide a foundation for subsequent research that relies heavily on the efficiency of technology in practice (e.g., counseling) and research to promote individuals' psychological well-being. Specifically, Connors and Rende's paper points to the potential of automatic coding to assess how individuals' physical and mental activities are related to subsequent decision-making through Movement Pattern Analysis. Zaytseva et al.'s paper discusses how establishing the link between errors (e.g., sequential and proximity) in cognitive test performance and brain networks may contribute to understanding the association between cognitive functioning and brain circuits and subsequent task performance among individuals with schizophrenia. This study, in particular, illustrates how psychological data science can pull together diverse types of data (e.g., cognitive tests and fMRI) to identify and understand the connections between brain networks and cognitive task performance to determine, as the authors' state, "the common denominators of the generalized deficits." Vitale et al. used a randomized crossover design to examine the effects of chronotype (i.e., circadian rhythmicity) on mood state (e.g., depression and anger) and perceived physical exertion following acute high intensity interval exercise. Based on prior research on the effects of online psychosocial intervention for engaging youth in mental health treatment, D'Alfonso et al. detail the development of the moderated online social therapy (MOST) project designed as an online peer support for youth experiencing mental health issues (e.g., depression). In particular, their work explores the ways in which computational and artificial intelligence methods (e.g., Chatbots) may offer a

mechanism to use automated user-specific therapy to supplement and enhance support delivered to patients by real-life moderators (and clinicians). Finally, Anwar et al. offer an informative book review on *Networks of the Brain*, authored by Olaf Sporns, which focuses on the application of network science in neuroanatomy. As presented in the review, the book offers several lines of inquiry for psychological data science regarding the data types that can contribute to modeling the brain within a complex network approach.

The papers included in this Research Topic illustrate the potential of psychological data science to unlock the wealth of information provided by diverse data types (e.g., physiological, social, and cognitive). The access to affordable, efficient processing is providing rich opportunities to implement a range of statistical and computational models and estimation procedures that were not readily available for use in applied research within the past few decades. Consequently, the rise of data science and big data is directly influencing the practice and research across a range of disciplines (e.g., marketing and medicine), including psychology. As such, the papers included in this Research Topic offer only a sampling of the body of work in the emerging discipline of psychological data science. Moving forward, we very much look forward to the ways in which data science and big data further develop the knowledge base in psychology.

AUTHOR CONTRIBUTIONS

JI and PC contributed to the writing of the editorial and review of the corresponding Research Topic papers.

Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health

Simon D'Alfonso [1,2]*, Olga Santesteban-Echarri [1,3,4], Simon Rice [1,3], Greg Wadley [2], Reeva Lederman [2], Christopher Miles [1], John Gleeson [5] and Mario Alvarez-Jimenez [1,3]

[1] Orygen, The National Centre of Excellence in Youth Mental Health, Melbourne, VIC, Australia, [2] School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia, [3] Centre for Youth Mental Health, The University of Melbourne, Melbourne, VIC, Australia, [4] Faculty of Education Sciences and Psychology, Universidad Rovira i Virgili, Tarragona, Spain, [5] School of Psychology, Australian Catholic University, Melbourne, VIC, Australia

Introduction: Benefits from mental health early interventions may not be sustained over time, and longer-term intervention programs may be required to maintain early clinical gains. However, due to the high intensity of face-to-face early intervention treatments, this may not be feasible. Adjunctive internet-based interventions specifically designed for youth may provide a cost-effective and engaging alternative to prevent loss of intervention benefits. However, until now online interventions have relied on human moderators to deliver therapeutic content. More sophisticated models responsive to user data are critical to inform tailored online therapy. Thus, integration of user experience with a sophisticated and cutting-edge technology to deliver content is necessary to redefine online interventions in youth mental health. This paper discusses the development of the moderated online social therapy (MOST) web application, which provides an interactive social media-based platform for recovery in mental health. We provide an overview of the system's main features and discus our current work regarding the incorporation of advanced computational and artificial intelligence methods to enhance user engagement and improve the discovery and delivery of therapy content.

Methods: Our case study is the ongoing Horyzons site (5-year randomized controlled trial for youth recovering from early psychosis), which is powered by MOST. We outline the motivation underlying the project and the web application's foundational features and interface. We discuss system innovations, including the incorporation of pertinent usage patterns as well as identifying certain limitations of the system. This leads to our current motivations and focus on using computational and artificial intelligence methods to enhance user engagement, and to further improve the system with novel mechanisms for the delivery of therapy content to users. In particular, we cover our usage of natural language analysis and chatbot technologies as strategies to tailor interventions and scale up the system.

Conclusions: To date, the innovative MOST system has demonstrated viability in a series of clinical research trials. Given the data-driven opportunities afforded by the software system, observed usage patterns, and the aim to deploy it on a greater scale, an important next step in its evolution is the incorporation of advanced and automated content delivery mechanisms.

Keywords: youth mental health, psychosis, depression, computational health, chatbots, sentiment analysis

INTRODUCTION

The majority of early intervention programs (specialist interventions and support to young people experiencing early symptoms of mental illness) offer services that do not last more than 2 years; for instance, early intervention services for psychosis typically offer support for 24 months. headspace, an Australian national foundation that provides early intervention mental health services to young people offers up to only 10 sessions of psychological therapy per year (Rickwood et al., 2014, 2015). As a result, some of the benefits gained through specialized treatment may not persist after its termination. Discharge and referral to a general mental health service may create a feeling of detachment among young people, decreasing engagement with mental health institutions and consequently increasing chances of relapse. In fact, reviews indicate that up to 80% of young people relapse from their initial condition after symptomatic remission from psychosis or depression (Alvarez-Jimenez et al., 2012).

Recent years have seen the development of online interventions to address these issues. Novel information and communication technologies provide an extraordinary opportunity for improving, and even transforming interventions with psychiatric disorders (Álvarez-Jiménez et al., 2016). Due to the general enthusiasm of young people for new technologies, with more than 97% of youth connecting to the Internet daily (Pew Research Center, 2014), internet-based interventions may be especially effective for, and attractive to, patients with different mental health disorders (Burns and Morey, 2008). Pioneering interventions using these technologies may play a pivotal role in addressing substantial challenges, comprising access to and engagement with services, and delivery of extended support to maintain the clinical gains of specialized services (Álvarez-Jiménez et al., 2012). Internet-based interventions may lead to the development of supportive relationships (O'Keeffe and Clarke-Pearson, 2011), decreased isolation (Dennis, 2003), increased self-disclosure (Weisband and Kiesler, 1996), and may possibly reduce stigma (Houston et al., 2002).

Social networking interventions in particular are uniquely placed to support young people experiencing mental ill health (Rice et al., 2014). Due to the stigma associated with mental illness, young people experience extreme social isolation and face difficulties in maintaining relationships (Morgan et al., 2011). The use of online social networking sites has been associated with positive socialization, promotion of supportive relationships, increased self-esteem, facilitation of communication, and feelings of group membership among young people, underscoring the relevancy among adolescents at risk of social isolation (Collin et al., 2011; O'Keeffe and Clarke-Pearson, 2011). These benefits may create a sense of belonging for young people, increasing use and engagement with social networks on the Internet.

The moderated online social therapy (MOST) project (Alvarez-Jimenez and Gleeson, 2012; Gleeson et al., 2012; Lederman et al., 2014) is designing, building and testing online social therapy systems for youth mental health. The MOST model uniquely integrates online peer support and evidence-based interventions with a clinician and consumer-centered service delivery process. It has also been developed following participatory design principles and it uses persuasive design elements to promote engagement with the intervention and behavioral change (Hagen et al., 2012).

To date, the MOST model has been effectively implemented in six studies, including four pilot studies; (i) Horyzons, for young people recovering from psychosis (Alvarez-Jimenez et al., 2013), (ii) Momentum, for young people at ultra-high risk of developing psychosis, (iii) Rebound, for young people recovering from depression (Rice et al., 2016), (iv) Meridian, for carers looking after young people experiencing mental health issues. In addition, there are currently two active longer-term randomized controlled trials evaluating MOST, (v) as a 5-year relapse prevention intervention for first episode psychosis (the Horyzons RCT), and (vi) a 2-year trial to support carers of young people with psychosis (the Altitudes RCT). Generation is also an upcoming trial in collaboration with eheadspace[1] (Rickwood et al., 2014, 2015) using the MOST software to power a general site for help seeking young people.

The MOST project was initiated with the aim of investigating two questions:

1. Efficacy of online therapy: do the clinical benefits of specialized face-to-face youth mental health programmes extend into long-term improvements through online psychosocial intervention?
2. Technology design: how can we best design and implement engaging technology in for young people with mental-ill health?

Recently a third question has arisen.

3. How can advanced computational and artificial intelligence (AI) methods be employed to supplement the support provided by moderators/clinicians and automate user-tailored therapy with a view to scaling up usage of the MOST model and platform?

Usage analyses of the MOST sites indicate that the social networking component is the most frequently used feature (Rice et al., 2016). Despite these favorable system usage statistics, we strive to improve the delivery of therapeutic aspects of the system to our users. Thus, this paper focuses on some recent research and incipient developments pertaining to the third question. We use the Horyzons site as our example, since it is the longest running implementation and has the largest set of usage data.

ACCESSING SITE CONTENT

A large focus of development efforts has been on ways in which online therapy content within MOST can best be delivered to users so that it is relevant, draws their interest and maximizes their engagement. MOST follows a positive psychotherapy model (i.e., strengths-based models; Seligman et al., 2006) and a theory-driven model of online human support by moderators (i.e., supportive accountability; Mohr

---

[1]https://www.eheadspace.org.au/

et al., 2011). The creation of therapy content in the MOST sites was driven by feedback from users and expert youth mental health clinicians through iterative prototyping and participatory design. The software system was designed via a tailored (bespoke) design process, which offered more flexibility to integrate social networking, therapy and the moderation component (Wadley et al., 2013). The result is a therapeutic environment where young people can learn and practice therapeutic techniques, gain perspective and validation, and learn how to solve problems in a transitional social network on their path toward recovery.

The modules were designed in a collaborative effort between professional writers for young people, clinicians, psychology researchers, and users (Lederman et al., 2014). Following is an outline of the MOST system's main parts:

1. *Social Networking and The Cafe*—A Facebook-style newsfeed where users can contribute posts and comments; share experiences, give and obtain support, and gain perspectives and validation. Additional features include a *Job Zone* for vocational opportunities and information and *Team Up*, where users can set a challenge for themselves that others can then also participate in or follow as part of a "cheer squad." See **Figure 1** for an illustration.

2. *Take a Step*—Steps are interactive therapy modules designed to exercise and develop a range of psychological skills. Therapy is delivered via engaging content that has been developed collaboratively by clinical psychologists, professional creative writers, leading comic developers, and young people. Social interaction is embedded within steps through *Talking Points*, which are questions that promote users to discuss and share their own experiences. See **Figures 2**, **3**, **4** for illustrations.

3. *Talk It Out*—A space where users can nominate problems or difficulties they would like some help with to discuss in moderated groups following an evidence-based social problem solving framework (McFarlane et al., 1995; McFarlane, 2004). Once a user has nominated a problem and framed it together with a moderator, solutions are proposed and discussed by the users before a moderator wraps up the thread with a synopsis. To date there are 75 completed Talk It Outs. Apart from the role each of these played in providing an active forum to address a user-nominated problem, the overall result is a valuable user generated "knowledge base" repository that can be searched and referred back to.

4. *Do It! (Actions)*—Users can access and "do" specific behavioral experiments or tasks (referred to as "actions") to apply mindfulness, self-compassion, and personal strengths in real world situations relevant to the young person (e.g., social context, school, work, alone, etc.). The use of context specific, action-based suggestions through online interventions has been recommended to change behavior, develop skills and increase practice, and generalization of these skills to real life situations (Van Gemert-Pijnen et al., 2014). Users can also bookmark actions so that they can easily return to and periodically reengage with those that they find helpful.

Conceptually, we can distinguish between The Cafe, a social space generated by user contributions and the Take a Step and Do It! sections, which offer authored therapeutic content for users to engage with. The Talk It Out section lies somewhere in between; Talk It Outs are generated by user contributions and the result is a valuable repository of material that users can access. Importantly, the MOST system also integrates these sections and functions; therapy content appears throughout the newsfeed and micro newsfeed-like discussion threads can occur in therapy pages. In general, integration is key and the system has been designed to create a constant back and forth flow for the user between therapy and social elements.

For example, suppose that a user starts on the newsfeed café page (**Figure 1**). They can comment on and react to posts from other users or they can add their own posts[2]. As will soon be covered in more detail, they also have access points to therapy content; moderator suggestions appear to the left, posts advertising content just taken by others users appear in the newsfeed and therapy suggestions relative to a user's own posts are made upon their submission. Suppose they click on the "How to flourish" step (**Figure 2**). While going through the step, they can interact with the step and contribute to a "mini-newsfeed" by commenting on a Talking Point (**Figure 3**). Upon completing the step, they can feed back into The Café by sharing a link to the step or rating the step and sharing this rating with an accompanying message. Also, a recommender system suggests other relevant steps the user might like to try (**Figure 4**).

DISCOVERING CONTENT

The most basic and direct ways in which users can access steps and actions are via a simple omnipresent search bar and the aforementioned primary navigation menu links. The ubiquity of search bars in this "Age of Search" make them a natural interface element familiar to Internet users. At present the search function accepts a simple search term and performs a basic text match against content in the system. Interestingly, it is not used a great deal and is used more so to search for users and site features rather than therapy content. We are currently looking into revamping this interface real-estate. Ultimately an "oracular" search box that performs a more sophisticated, customized form of information retrieval could process expressive input from the user and instantly provide a gateway to relevant therapy content.

Under the steps section users can browse an alphabetically-ordered grid of all the steps presented in a visually appealing manner with respective icons, as illustrated **Figure 5**.

Given the potential for users to be encouraged to participate based on the activity of their peers, we also include a page "Steps People are Taking." The four access links featured on this page are:

1. Steps taken recently—Among all the steps visited from these four options, 48% of the visits resulted from this link.

---

[2]The posts shown in **Figure 1** are actually real samples taken from the live Horyzons system.

FIGURE 1 | The Horyzons Cafe newsfeed.
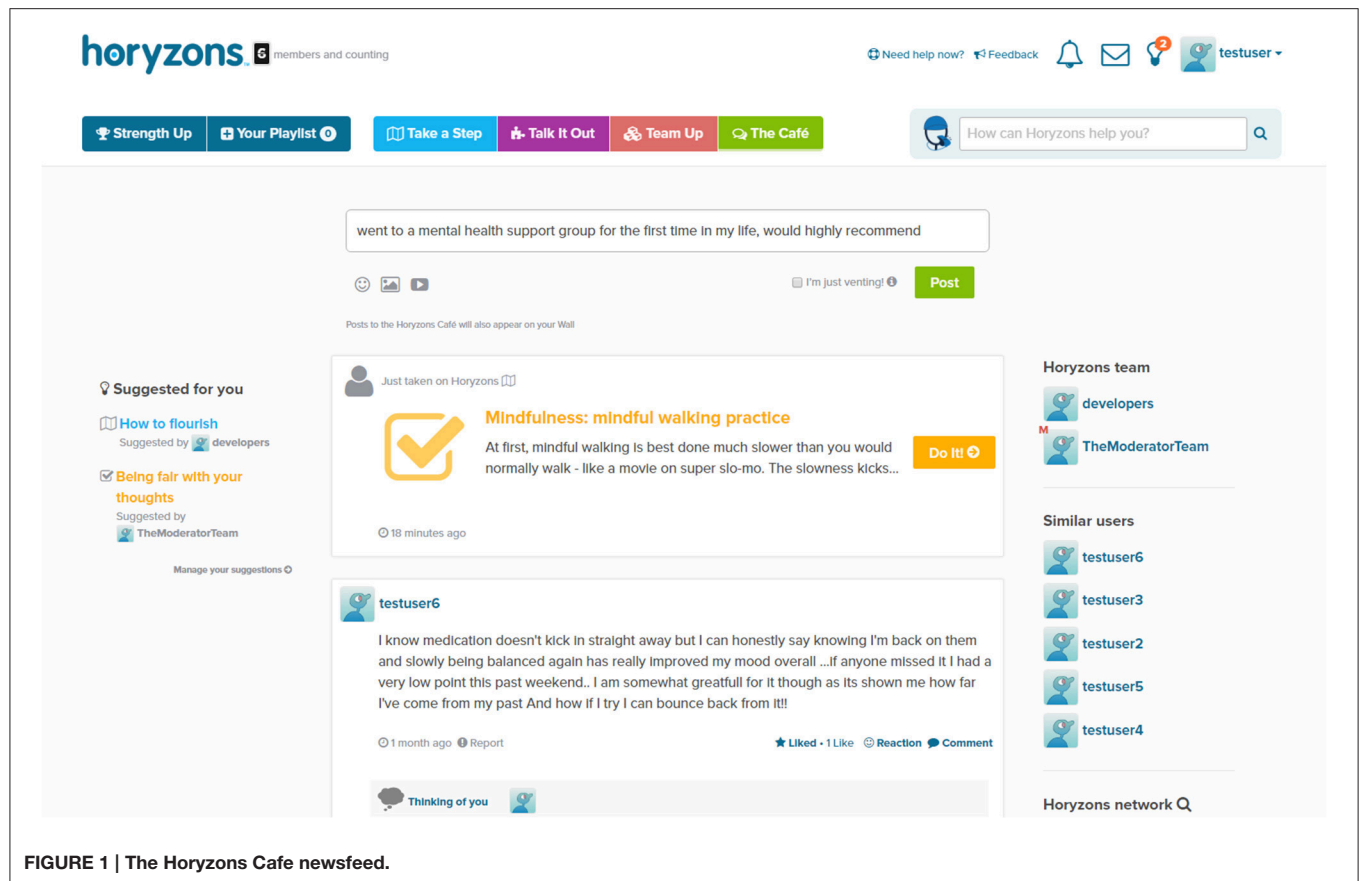
2. Steps that have got people talking—Users are presented with the opportunity to comment on Talking Points at various points during steps. This option selects those steps with the most Talking Point comment activity. Accounts for 14% of all step visits.
3. Steps taken most often—Accounts for 11% of all step visits.
4. Hidden Treasures—A modest means of countering informational cascades (Easley and Kleinberg, 2010) and promoting less popular steps by displaying those steps that have been done the least. The euphemism seems to have been effective, as this link accounts for 27% of all step visits from these four options.

Actions have a similar interface and users can also engage with actions in the "Powered by your Strengths" page. Upon joining the system, users can complete an initial exercise where they choose 5 out of 24 strengths that they believe best apply to them. Some example strengths are Courage, Discretion, Creativity, and Curiosity. Strengths are connected to relevant actions through which they can be exercised and this option simply presents users with actions that are connected to the strengths they have chosen.

Whilst the presence of these structures provides a straightforward means to access content, their supplementation with more advanced, automated and user-tailored options is desirable. The following considerations concerning the limitations of these standard access options serve as a starting

point for the justification of such supplementation. Search bars are a simple, effective means to find content. Sometimes though a user might not have the right search term in mind for what they are seeking or they might not have the intention to seek. Also, it could be that their behavior on the site suggests that they could benefit from something they are not even aware of. Direct menu links are good for users who are keen to browse the therapy content or for those who have the motivation to search for something they have in mind. This will often not be the case though and we cannot always expect this from users. Thus, rather than expecting users to always seek content or know what they want to search for, we endeavor to develop more sophisticated forms of content delivery. As a starting point for this endeavor, we believe that users may benefit from more advanced taxonomic guidance in finding relevant content. This consideration led to the incorporation of a tag-based system.

THERAPY TAGS

The purpose of tags is to create meaningful categories based on therapeutic content, which addresses target groups of symptoms (i.e., anxiety) or skills (i.e., social abilities). This classification allows both users and clinicians to find or suggest individually-tailored content for each person, taking into account the particular emotional state and their needs at that precise moment.
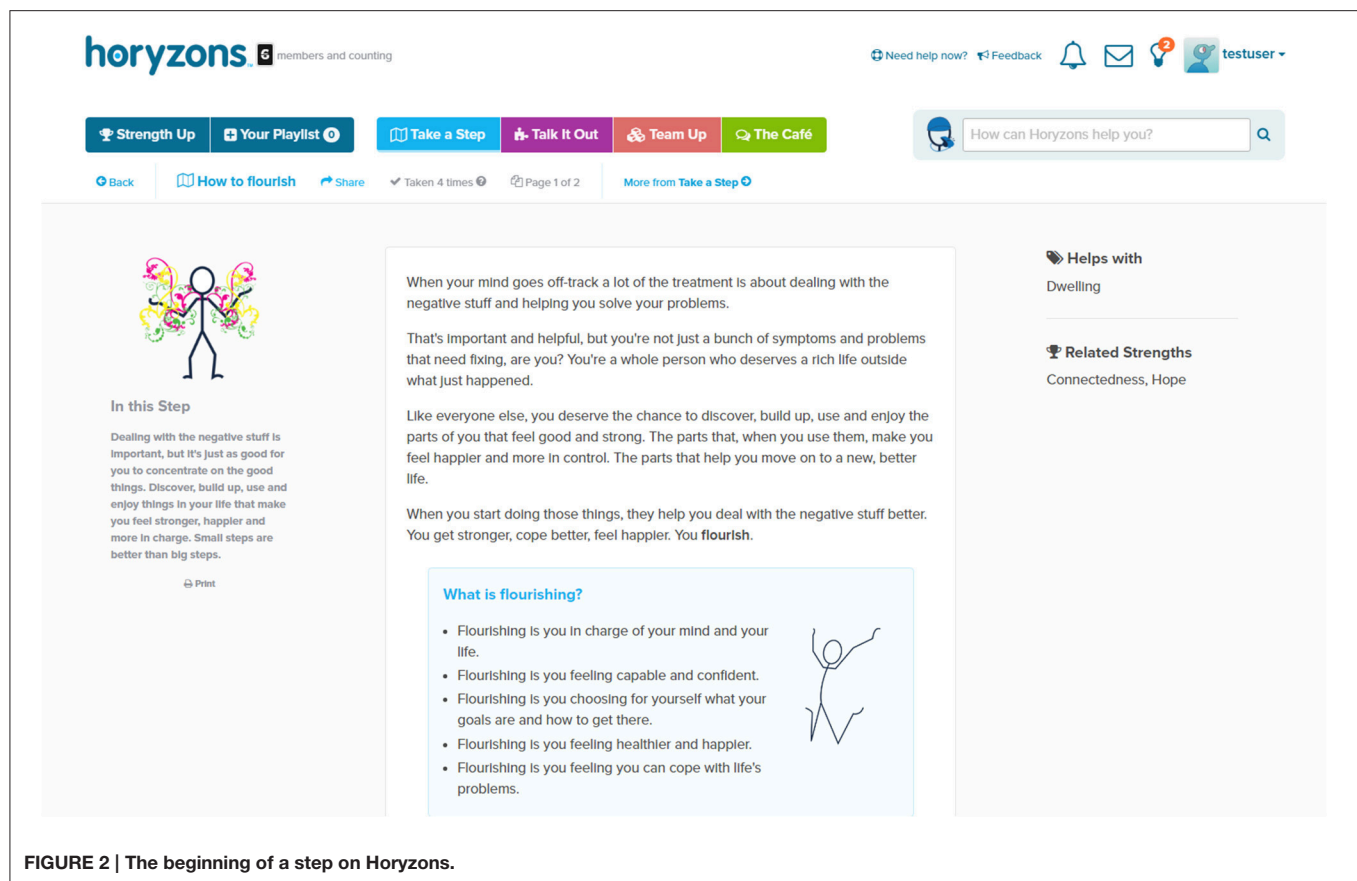
FIGURE 2 | The beginning of a step on Horyzons.

There are three parent tags in the system and each of these has three sub-tags:

- Coping
  - Solving Problems
  - Kicking the Habit
  - Beating Sadness and Worries

- Connecting
  - Overcoming Conflict
  - Boosting Relationships
  - Stories Like Yours

- Enhancing
  - Making Happiness
  - Wellbeing with Mindfulness
  - Work and Study

Steps and actions are tagged with one or more of these sub-tags and users can accordingly view or search for this content via these tags.

HUMAN-SUPPORTED ENGAGEMENT

Apart from providing navigational/taxonomic structures to access therapy content, the MOST system also involves the direct delivery of specific content suggestions to individual users. From its inception, a feature of the MOST system has been that moderators can select content to suggest to users. Moderation by staff has been essential to increase adherence, because positive user motivation is enhanced by accountable and trusted experts who moderate site usage (Mohr et al., 2011). Moderators not only provide safety by preventing misuse of the system, but encourage usage and enhance user experience by acting as a role-model. The moderation team includes seven clinical psychologists and a clinical social worker. Moderation of specific topics is provided by a vocational worker and an expert in youth participation. Moderator therapy suggestions are based on what the user's moderator deems to be appropriate based on knowledge of factors such as the user's history and profile, as well as their current engagement with and activity on the site. For example, one of our moderators was informed via discussion with a user that the user had recently gained employment. With knowledge of both this and the fact that the user had not yet chosen any strengths, the moderator suggested the step "*Strengths for work*," which helps users to find ways to use their strengths to survive and thrive at work. If a suggested content item is completed by the user, then the system automatically records this. There are also the options of a client or a moderator dismissing the suggestion. Naturally, such information helps to inform subsequent moderator suggestion choices.
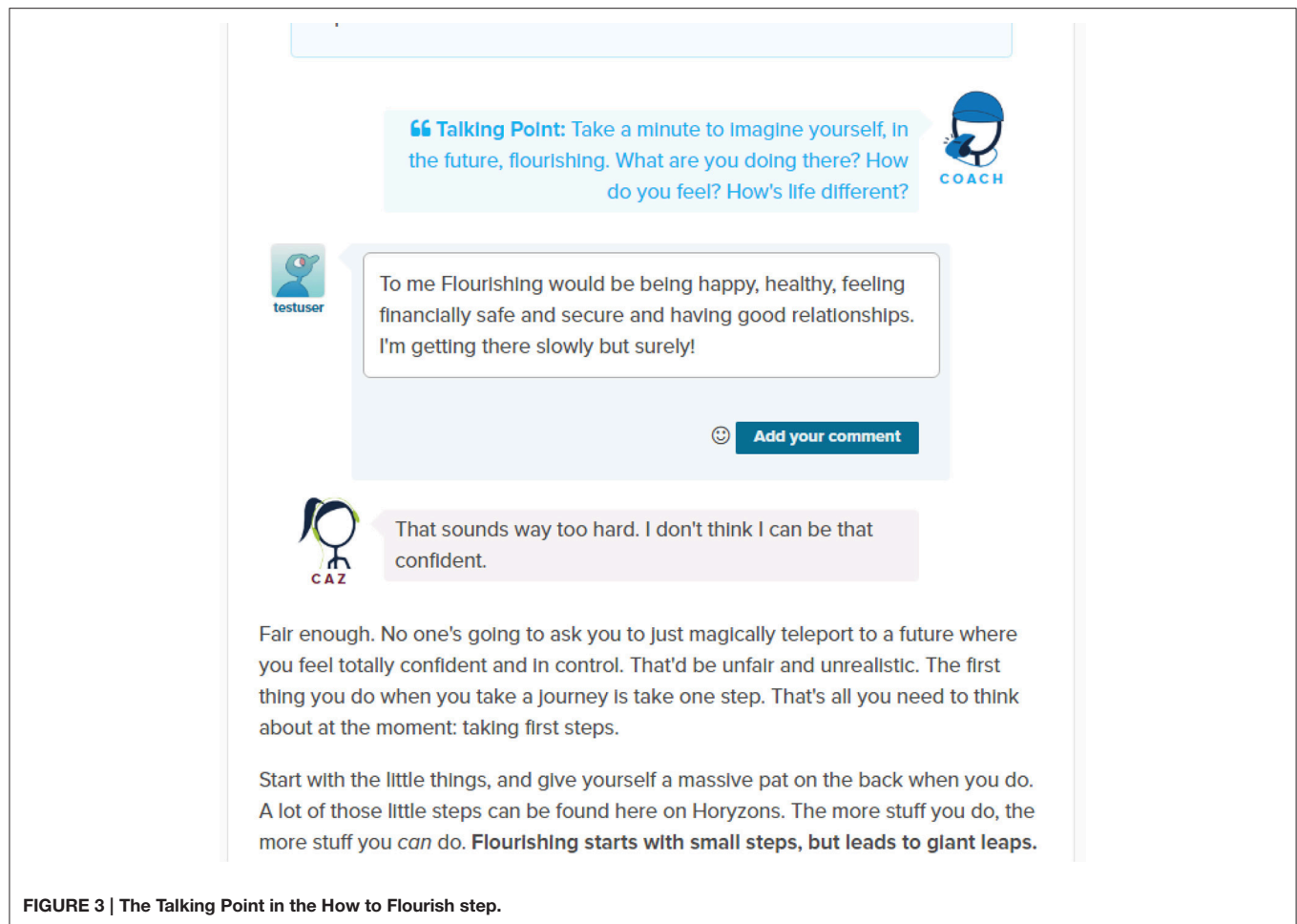
FIGURE 3 | The Talking Point in the How to Flourish step.

The main advantage of moderator suggestions is that they are customized based on expert moderator assessments of an individual user and what therapy content is considered appropriate for them. The personalized suggestion perhaps makes the user more receptive to engagement with the content. To date, there have been a total of 701 moderator-to-user suggestions. Out of this number, 211 have been completed (21 have been client dismissed, 304 moderator dismissed, and 165 pending). This is actually a fairly successful strike rate relative to other therapy content access points in the system. Our content completion rates compare well with short-term Internet-based programs (Christensen et al., 2004), and also compared to online interventions comprising some kind of moderation by interviewers or counselors (Christensen et al., 2004; Clarke et al., 2005). Furthermore, of the various therapy content access points, moderator suggestions have the highest usage number; around 39% of all tracked visits to steps/actions are due to moderator suggestions.

AUTOMATED SUGGESTIONS

Whilst manual moderator suggestions are a key part of our current sites, we are investigating automated methods for content delivery and how they can be used to facilitate user engagement; complementing our current models and addressing limitations. The first obvious advantage of automated suggestions over moderator suggestions is that they are not limited to times when moderators are available on the system and furthermore they can be delivered to the user in real-time. Automated suggestions also facilitate scaling the site to a larger user group. At present, the MOST platform is used with relatively small groups of users in a research setting (numbers ranging between 30 and 100 as opposed to hundreds, thousands, or more). Each moderator is assigned a manageable list of users, which they monitor and attend to. Although, such automated therapy suggestion methods currently serve as an alternative rather than a replacement for moderator suggestions, given the aim of scaling up and deploying the MOST platform in a more general, less moderated, or even unmoderated publicly available setting, automation becomes a prime goal.

Originally, the only form of automated content suggestion occurred when a user did a step. In a manner similar to Amazon's related recommendations, upon completing a step, a list of related steps is dynamically presented to the user, preceded by the message "*You might also be interested in.*" Around 5% of all tracked step visits are from this path. Related steps are
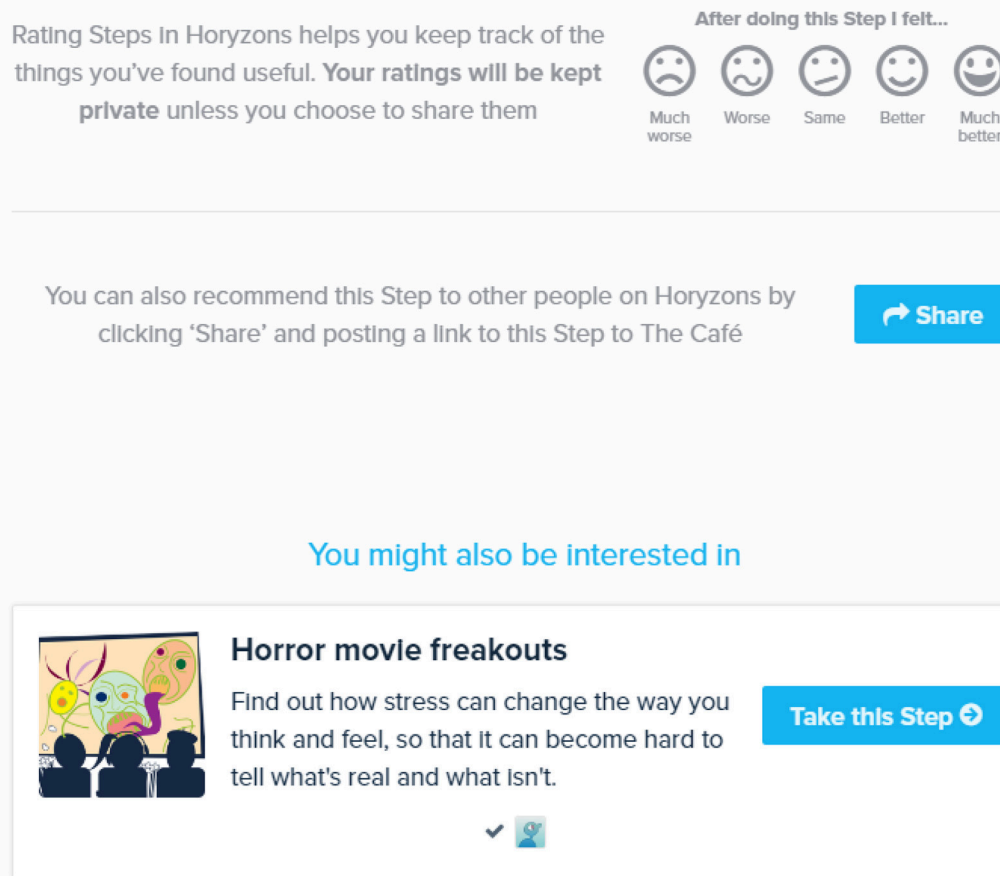
FIGURE 4 | Finishing the step How to Flourish.

determined by step-to-step connections pre-set by the system author. The next subsequent additions involved inserting therapy suggestions into the newsfeed. Firstly, when a step/action is done, this fact is anonymously posted in the newsfeed. Secondly, we make select action suggestions into the newsfeed, in a manner similar to Twitter/Facebook feed ad placements. These suggestions are based on the user's chosen strengths and any steps they have taken. Around 9% of all tracked action/step visits are from this path.

These two features offer an easy to implement, cost-effective way to promote therapy content compared to the suggestions made by a moderator. Human presence is not needed, the user receives the suggestion immediately, and if they perform another activity, a new related suggestion will appear in their screen. However, their delivery is not necessarily user-tailored or immediately relevant. A recently incorporated feature addresses such issues and connects newsfeed posting activity with therapy content suggestions. The basic idea is that linguistic analysis

of user posts can extract certain information on which to base content suggestion. As soon as a user submits a post, an algorithm sets off to analyse the post and determine one step, and one action, that are relevant to the post. The way these suggestions are presented has raised some usability issues and questions. Due to the calculations and remote application programing interface (API) calls (3Scale NSL, 2011) required, it can take up to 10 s in order to calculate these suggestions. This lag initially posed a significant problem from a user interaction perspective, as the calculation was made sequentially before the post was added to the newsfeed, leaving the users wondering what was happening. We therefore decided to calculate the suggestions parallel to the posting of the post. Upon addition of a post to the newsfeed, users receive the message "*Horyzons has suggestions based on your post*" above the post. If the user presses the "*Show Me*" button next to this message, the system attempts to retrieve the suggestions and display them above the post. If the suggestions have not been calculated yet, then the system
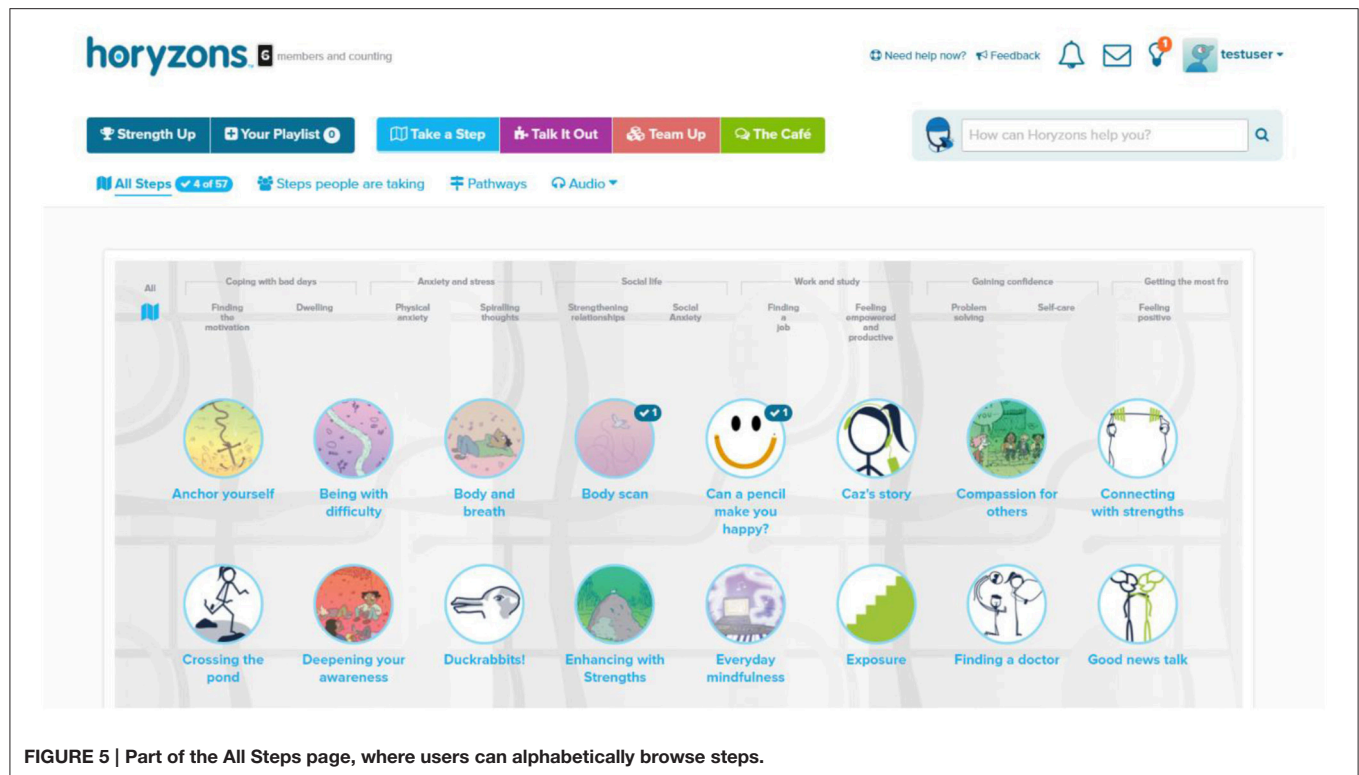
FIGURE 5 | Part of the All Steps page, where users can alphabetically browse steps.

displays a "*Delivering your suggestions*" message along with a dynamic progress image. It is hoped that this provides users with a sense that the calculation is very much occurring and it is occurring specifically in response to their posting activity. In order to foster a sense of privacy and individuality, we also accompany this with the message "These suggestions are only visible to you."

With this suggestion delivery interface now determined, our focus is on developing and experimenting with the underlying algorithm that calculates the suggestion. At present, we use a combination of the post's sentiment, emotion scores, and keywords to determine its most relevant steps and actions. For the reader's interest, a brief description of our current algorithm is as follows:

1. Upon submission of a post to the newsfeed, a call is made to IBM's Alchemy [10] text analysis system. First, it provides a score for the post's sentiment, with score $= 0$ being neutral, $1 >$ score > 0 being positive, and score $0 >$ score > -1 being negative. Second, it provides a score between 0 and 1 for the emotions of Anger, Disgust, Fear, Joy, and Sadness. Third, it extracts the post's keywords[3].

2. Our steps and actions have been partitioned into those which are suitable for positive posts, those which are suitable for negative posts and those which are suitable for either. We use this partitioning such that if a post is positive then a set of suggestion candidate steps/actions is reduced to those marked positive. Similarly for negative. If the post is neutral, then no such reduction is made.

3. Our steps and actions have also been assigned measures for each of the five emotions mentioned in point 1; the higher the measure, the more relevant the item is for posts exhibiting that emotion. We also have extracted and stored the keywords for all of our steps and actions. Given this, each of the candidate steps/actions is assigned a score based on their semantic similarity with the post's keywords and congruence with the post's emotion measures.

4. The step with the highest score is selected as is the action with the highest score. If more than one candidate emerges, at this stage one is randomly chosen.

The keyword semantic similarity component is the most decisive factor in these calculations. Given that much of our therapy content is specifically suited to only positive or negative states, the sentiment analysis component determines an important initial partitioning of therapy content where applicable. The emotion congruence component provides a balancing and adjudicating factor. First, it can distinguish therapy items that have the same semantic similarity score. Second, it can "downgrade" therapy items with a relatively high semantic similarity score that are otherwise emotionally inappropriate relative to the post.

---

[3]The IBM Alchemy service has been employed as it offers affordable expert text analysis capabilities accessibly, allowing us to straightforwardly implement our proof of concept. Limitations regarding its proprietary and unscrutinised nature will be something to consider. Short of conducting our own program into text analysis, academically reviewed alternatives include Valence Aware Dictionary and sEntiment Reasoner (VADER), Linguistic Inquiry and Word Count (LIWC), and SentiStrength. One advantage of some of these alternatives is that the software libraries could be hosted on our own servers.

Figure 6 illustrates a simple example post and accompanying suggestions. As can be gathered from their titles, the suggested step and action are quite relevant relative to the post content.

Once the page is reloaded these suggestions disappear. As **Figure 7** illustrates, they do though remain visible to moderators, along with sentiment and emotion information.

This initial algorithm has just been implemented and we will soon be analyzing our records of suggestions that have been made in order to gain insight into how it can be improved. Ultimately, there are many linguistic/psychometric properties that could be of use in pairing posts with relevant therapy content (Tausczik and Pennebaker, 2010).

There are also plans to include a feedback interface next to the suggestions so that users can rate how relevant the suggestions were for them. This data will help to influence our algorithm and content scoring in general and will also provide us with information on how we can further tailor content to specific users. Apart from refining the custom selection of content suggestions, there is also the possibility of tailoring the message given to a user upon suggestion of a piece of therapy content. By analyzing a post and other historical information about the user, we could offer messages of the form "This step was recommended to you because . . . ." Beyond using and responding to single posts, there is also the possibility of using larger segments of user input accumulated over time. Along with other pertinent user usage data, these segments could help to determine periodic automated suggestions akin to moderator suggestions that are delivered independently of and not in response to a specific user action on the site. Finally, it should be noted that statistics provided over the last few sections are used to make comparisons between the ways in which site content can be accessed; how people are navigating the site and their preferred methods for accessing content. We are yet to conduct any A/B or split testing to evaluate the impact of site features, although this is something that is on our agenda. Before closing, we will next briefly look at the possibility of incorporating chatbots into online social therapy.

CHATBOTS

A chatbot is a computer program that mimics conversation with users via a chat interface, either text or voice based (Abdul-Kader and Woods, 2015). The underlying system can be based on a variety of foundations, ranging from a set of simple rule-based responses and keyword matching to powerful natural language processing (NLP) (Chowdhury, 2003) and machine learning (ML) (Smola and Vishwanathan, 2008) techniques; NLP concerns the use of computers to understand and manipulate natural language and ML concerns self-learning computer programs with the ability to grow and adapt in response to new data, without being explicitly programmed to do so (Saeed, 2016). Irrespective of the actual intelligence of the responding bot, there is something distinct about the experience of a user entering input and a bot responding. A bot may create the sensation of a natural and real environment for the user since it can understand natural speech patterns. While an app or a web search may give the user a direct

answer in response to a search query, a bot simulates a real-life conversation as if the user was talking with another person; the uniqueness of the feature resides in the user's perception of an interaction (Margalit, 2016). Also, beyond the challenge of imbuing chatbots with intelligence in terms of their ability to simulate the structures of natural language communication, another important dimension, particularly in a psychology/therapy setting, is emotional intelligence; developing chatbots that can detect and respond appropriately to the emotional state of the human. Some recent work on emotionally intelligent AI comes from the province of affective computing[4] (Brewster, 2016; Skowron et al., 2017).

Chatbots are currently a hot topic in the tech world, with major technology companies such as Facebook, Microsoft, and Google making significant investment forays into this emerging technology. The commercial applications of chatbots range from the provision of online customer service to conversation-based product searches and event organization. Whilst our motivations differ, this current commercial interest in chatbots makes for a particularly opportune time to consider their incorporation into online mental health platforms such as MOST. In fact, the history of chatbots is intimately tied with psychology. Apart from the interesting philosophical and psychological questions they raise, the first well-established chatbot, ELIZA was actually programmed (in 1966) to simulate a Rogerian psychotherapist (Weizenbaum, 1966). A modern ELIZA, the more sophisticated Artificial Linguistic Internet Computer Entity (A.L.I.C.E.), first surfaced in 1995 (Shah, 2006). Its creation resulted in the development of a general usage Artificial Intelligence Markup Language (AIML), which can be used to quickly create a basic bot from scratch. ALICE is a thrice winner of the Loebner Prize, an annual competition in artificial intelligence that awards chatbots judged to be the most human-like. The format of the competition is based on a standard Turing test (Shah, 2006).

Whilst it could be said that the ultimate goal regarding chatbots is to truly pass the Turing test (Turing, 1950) and convince a human judge that it (the bot) is human (Oppy and Dowe, 2016), our present focus is on using chatbots as conversational search/assistant interfaces. Instead of finding therapy content via the search box or menus, a chatbot offers an alternative search modality. At its simplest, a chatbot could guide the user in terms of disclosing emotions, therapy preferences and needs. Upon receiving the user's input, the bot could respond with some suggestions. User experience elements are critical as there are certainly cases where standard search interfaces/apps are a better, more effective choice than chatbots (Verber, 2016). There is however distinct potential in using chatbot searches. A conversational search mode can create a sense of connectivity and personalization that offers a uniquely effective way to collect input from users. Research suggests that users are more open and likely to share information, particularly on sensitive topics, when interacting with a machine interface (Weisband and Kiesler, 1996; Tantam, 2006; Lucas et al., 2014), thus it

---

[4]http://www.affectiva.com/
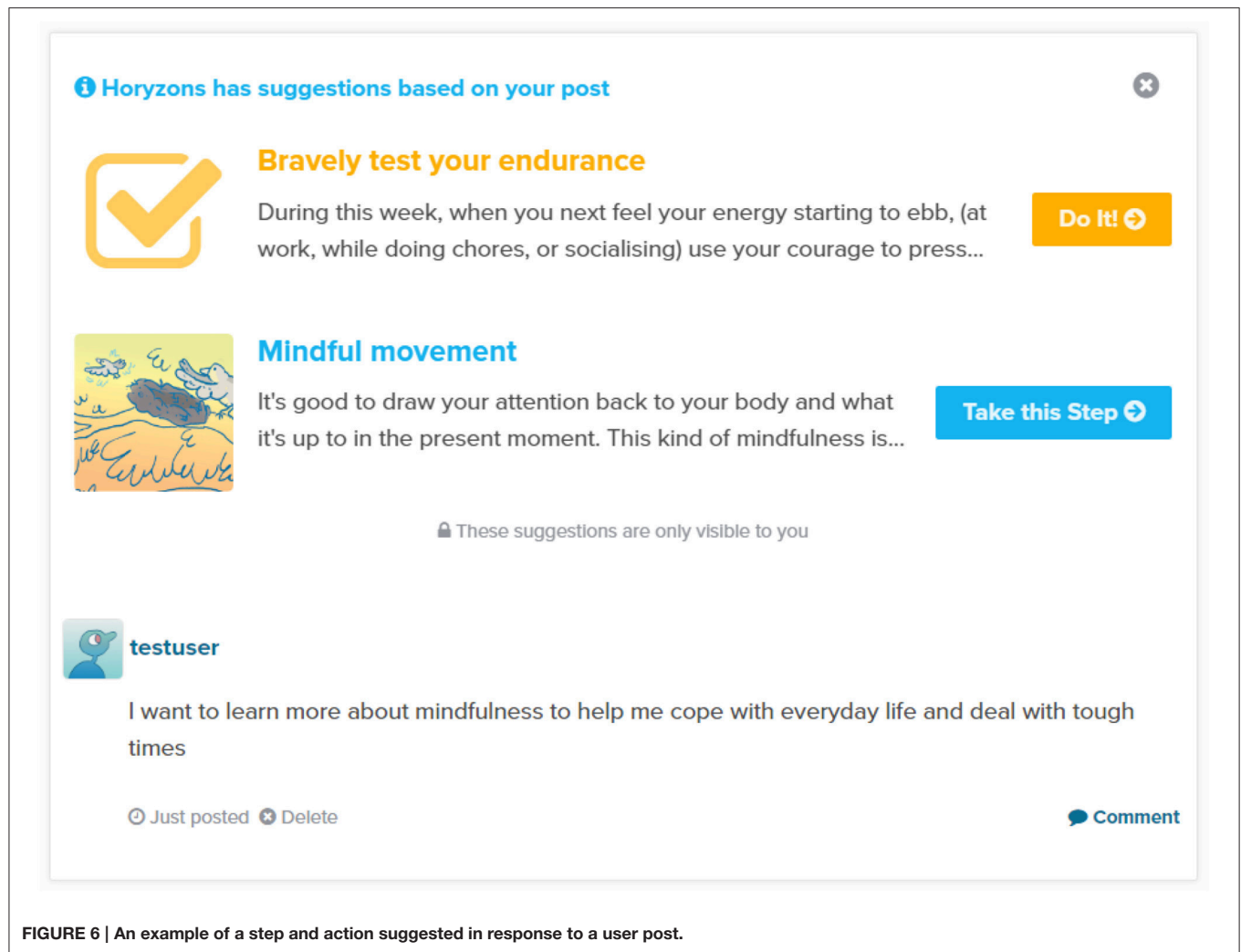
FIGURE 6 | An example of a step and action suggested in response to a user post.
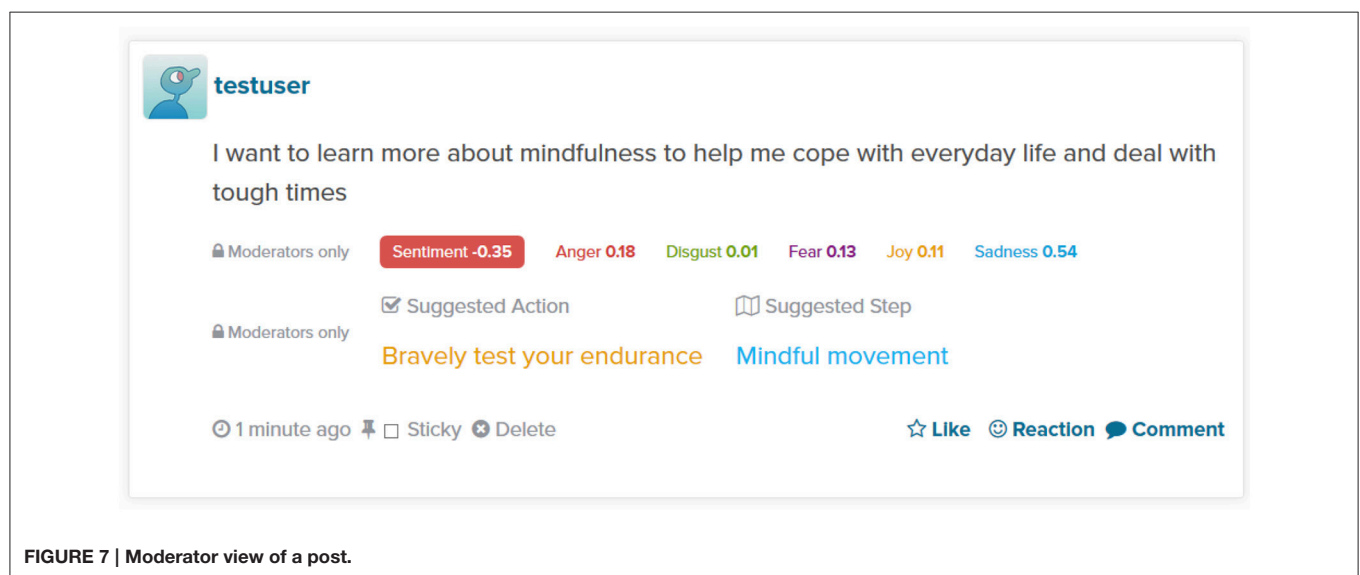
FIGURE 7 | Moderator view of a post.

may be important to include this technology in mental health web-sites due to the stigmatizing and delicate nature of the topic.

Transforming search from a user-initiated task to a quasi-conversation can be more conducive to eliciting information from users regarding their wants and needs. Furthermore, chatbots could perform psychological assessment in a novel way and learn from user responses in real time. We could also tailor the way the suggestion is made relative to the user, as mentioned in the previous section. One other interesting aspect to consider is the extra secondary information that can be obtained from even a few input sentences from the user. As opposed to a static form with pre-defined responses or a search text box that receives one or two word inputs, the input gathered from a conversational search is going to have a richer structure. Linguistic analysis of this input could then contribute to assessing a user's immediate requirements as well as forming a part of an overall assessment of the user.

Beyond this, there are some other interesting usage possibilities in between a basic conversational search and a chatbot that is the equivalent of a human mental health professional. Whilst not yet being sophisticated enough to replicate a therapist, bots that can maintain a basic form of conversation beyond one-question, one-input, one-response are feasible and have been implemented for different purposes: as a virtual dietician for diabetic patients (Lokman and Zain, 2009); as an educational system for students (Mikic et al., 2009); and as an E-learning system for speech learning for disabled people (Bhargava and Nikhil, 2009) among others. The implementation of such a chatbot would mean that more conversational text could be collected from the user. A richer body of conversation to analyse presumably translates to the possibility of better content suggestions.

Also, more sophisticated conversational bots like this could be used to gather information about a user before they chat online with a real therapist. Whilst not yet activated on any of our trials, the MOST system does have a real-time client to human moderator online text chat feature. Upon requesting to chat, users are asked to complete a number of questions before being sent to a waiting queue until a moderator is ready to accept their request. The information that they fill in helps the moderator to assess the client and their priority. Rather than just being restricted to a set of rigid form questions, a chatbot could be used to converse with the client before the human moderator is available to accept their request. Upon doing so, the bot conversation text can be analyzed and used to provide the moderator with a pre-profile of the client with whom they are about to chat.

CONCLUSIONS AND FUTURE WORK

The innovative MOST system, which integrates online peer support and evidence-based interventions with a unique clinician and consumer-centered service delivery process, has thus far demonstrated its viability in a series of research trials. Given the data-driven opportunities afforded by the system and the

aim to deploy it on a greater scale, an important next step in its evolution is the incorporation of advanced computational and artificial intelligence methods. This innovation will play a central role in furthering foundational goals of our interventions such as enhancing user engagement, facilitating the discovery and delivery of tailored therapy content and promoting autonomy, competence and relatedness. It will also serve to supplement, complement, or possibly even replace human moderators. Ultimately this endeavor will lead to a more scalable system that is better positioned to meet the unmet needs in large-scale mental health provision and long-term support.

A current focus of our work is on developing mechanisms to deliver individualized therapy suggestions based on linguistic analysis of newsfeed postings and other pertinent factors such as user preferences and histories. This is one specific example of how the analysis of user content can feed information retrieval services and is related to the more general field of using computational linguistic analysis to predict/determine psychological states and characteristics (Tausczik and Pennebaker, 2010; Gkotsis et al., 2016). As we move from controlled trials with a select group of participants each having the same established mental health condition to more general, publicly used sites where users will have a variety of conditions and are not pre-known trial participants, there is also the potential to deliver therapy content particularly relevant to a certain mental health condition based on analysis of a user's content (Bedi et al., 2015).

Apart from user content analysis, other possibilities include making suggestions based on geolocation data (Dredze et al., 2013) and self-monitoring/self-sensing data (Matthews et al., 2014). Regarding the former, one example would be to determine the geographical location or event of an active user using their mobile phone, retrieve an action that could make use of that location/event and deliver the suggestion via a mobile notification. Regarding the latter, one example would be to detect the onset of anxiety attacks with wrist sensor technology (Talbot, 2012; Kappas et al., 2013) and provide the user with appropriate therapy content. As discussed, we have also done some preliminary exploration into the implementation of therapy chatbots, which can simulate interactions with human beings to varying degrees of reality and will quite possibly become a mainstay of future e-mental health and help seeking web sites/applications.

AUTHOR CONTRIBUTIONS

REFERENCES

3Scale NSL (2011). *What is an API? Your Guide to the Internet Business (R)evolution*. 3 Scale Infrastruct Program Web, 1–9.

Abdul-Kader, S. A., and Woods, J. (2015). Survey on chatbot design techniques in speech conversation systems. *Int. J. Adv. Comput. Sci. Appl.* 6, 72–80. doi: 10.14569/IJACSA.2015.060712

Alvarez-Jimenez, M., Bendall, S., Lederman, R., Wadley, G., Chinnery, G., Vargas, S., et al. (2013). On the HORYZON: moderated online social therapy for long-term recovery in first episode psychosis. *Schizophr Res.* 143, 143–149. doi: 10.1016/j.schres.2012.10.009

Alvarez-Jimenez, M., and Gleeson, J. F. (2012). Connecting the dots: twenty-first century technologies to tackle twenty-first century challenges in early intervention. *Aust. N.Z. J. Psychiatry* 46, 1194–1195. doi: 10.1177/0004867412464067

Álvarez-Jiménez, M., Gleeson, J. F., Bendall, S., Lederman, R., Wadley, G., Killackey, E., et al. (2012). Internet-based interventions for psychosis: a sneak-peek into the future. *Psychiatr. Clin. North Am.* 35, 735–747. doi: 10.1016/j.psc.2012.06.011

Álvarez-Jiménez, M., Gleeson, J. F., Rice, S., Gonzalez-Blanch, C., and Bendall, S. (2016). Online peer-to-peer support in youth mental health: seizing the opportunity. *Epidemiol. Psychiatr. Sci.* 25, 123–126. doi: 10.1017/s2045796015001092

Alvarez-Jimenez, M., Priede, A., Hetrick, S. E., Bendall, S., Killackey, E., Parker, A. G., et al. (2012). Risk factors for relapse following treatment for first episode psychosis: A systematic review and meta-analysis of longitudinal studies. *Schizophr Res.* 139, 116–128. doi: 10.1016/j.schres.2012.05.007

Bedi, G., Carrillo, F., Cecchi, G., Slezak, D. F., Sigman, M., Mota, N. B., et al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia* 1, 1–7. doi: 10.1038/npjschz.2015.30

Bhargava, V., and Nikhil, M. (2009). An intelligent speech recognition system for education system. Available online at: https://pdfs.semanticscholar.org/8ea7/725bab4e390e4d8ab8ebee747d2a5340c3b2.pdf

Brewster, S. (2016). Customer service bots are getting better at detecting your agitation. *MIT Technol. Rev.* Available online at: https://www.technologyreview.com/s/602352/customer-service-bots-are-getting-better-at-detecting-your-agitation/

Burns, J., and Morey, C. (2008). "Technology and young people's mental health and wellbeing," in *Challenges in Adolescent Health: An Australian Perspective*, eds D. L. Bennet, S. J. Towns, E. J. Elliot, J. Merrick (Melbourne, VIC: International Academic Press), 61–71.

Chowdhury, G. (2003). Natural language processing. *Annu. Rev. Inf. Sci. Technol.* 37, 51–89. doi: 10.1002/aris.1440370103

Christensen, H., Griffiths, K. M., Korten, A. E., Brittliffe, K., and Groves, C. (2004). A comparison of changes in anxiety and depression symptoms of spontaneous users and trial participants of a cognitive behavior therapy website. *J. Med. Internet Res.* 6, 1–11. doi: 10.2196/jmir.6.4.e46

Clarke, G., Eubanks, D., Reid, E., Kelleher, C., O'Connor, E., DeBar, L. L., et al. (2005). Overcoming depression on the internet (ODIN) (2): a randomized trial of a self-help depression skills program with reminders. *J. Med. Internet Res.* 7, 1–12. doi: 10.2196/jmir.7.2.e16

Collin, P., Rahilly, K., Richardson, I., and Third, A. (2011). *Literature Review: The Benefits of Social Networking Services*. Melbourne, VIC. Available online at: http://www.fya.org.au/app/theme/default/design/assets/publications/

Dennis, C.-L. (2003). Peer support within a health care context: a concept analysis. *Int. J. Nurs. Stud.* 40, 321–332. doi: 10.1016/S0020-7489(02)00092-5

Dredze, M., Paul, M., Bergsma, S., and Tran, H. (2013). "Carmen: a twitter geolocation system with applications to public health," in *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*. Available online at: http://www.aaai.org/ocs/index.php/WS/AAAIW13/paper/download/7085/6497

Easley, D., and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, NY: Cambridge University Press.

Gkotsis, G., Oellrich, A., Hubbard, T. J. P., Dobson, R. J. B., Liakata, M., Velupillai, S., et al. (2016). "The language of mental health problems in social media," in *Proceedings of the 3rd Work Computational Linguistics Clinical Psychology from Linguist Signal to Clinical Reality* (San Diego, CA), 63–73.

Gleeson, J. F., Álvarez-Jiménez, M., and Lederman, R. (2012). Moderated online social therapy for recovery from early psychosis. *Psychiatr. Serv.* 63:719. doi: 10.1176/appi.ps.20120p719

Hagen, P., Collin, P., and Metcalf, A. (2012). *Participatory Design of Evidence-Based Online Youth Mental Health Promotion, Intervention and Treatment*. Melbourne, VIC: Young and Well Cooperative Research Centre.

Houston, T. K., Houston, T. K., Cooper, L. A., Cooper, L. A., and Ford, D. E. (2002). Internet support groups for depression: a 1-year prospective cohort study. *Am. J. Psychiatry* 159, 2062–2068. doi: 10.1176/appi.ajp.159.12.2062

Kappas, A., Küster, D., Basedow, C., and Dente, P. (2013). "A validation study of the Affectiva Q-Sensor in different social laboratory situations," in *53rd Annual Meeting of the Society for Psychophysiological Research* (Florence).

Lederman, R., Wadley, G., Gleeson, J. F., Bendall, S., and Alvarez-Jimenez, M. (2014). Moderated online social therapy: designing and evaluating. *ACM Trans. Comput. Interact.* 2, 1–27. doi: 10.1145/2513179

Lokman, A. S., and Zain, J. M. (2009). "An architectural design of virtual dietitian (ViDi) for diabetic patients," in *2nd IEEE International Conference on Computer Science and Information Technology* (Beijing), 408–411. doi: 10.1109/iccsit.2009.5234671

Lucas, G. M., Gratch, J., King, A., and Morency, L. P. (2014). It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* 37, 94–100. doi: 10.1016/j.chb.2014.04.043

Margalit, L. (2016). *The Psychology of Chatbots*. (Accessed August 19, 2016). Available online at: https://www.psychologytoday.com/blog/behind-online-behavior/201607/the-psychology-chatbots

Matthews, M., Abdullah, S., Gay, G., and Choudhury, T. (2014). Tracking mental well-being: balancing rich sensing and patient needs. *Computer (Long Beach, CA).* 47, 36–43.

McFarlane, W. R. (2004). *Multifamily Groups in the Treatment of Severe Psychiatric Disorders*. New York, NY: Guilford Press.

McFarlane, W. R., Lukens, E., Link, B., Dushay, R., Deakins, S. A., Newmark, M., et al. (1995). Multiple-family groups and psychoeducation in the treatment of schizophrenia. *Arch. Gen. Psychiatry* 52, 679–687. doi: 10.1001/archpsyc.1995.03950200069016

Mikic, F. A., Burguillo, J. C., Llamas, M., Rodríguez, D. A., and Rodríguez, E. (2009). "CHARLIE: an AIML-based chatterbot which works as an interface among INES and humans," in *IEEE Xplore Conference: EAEEIE Annual Conference*.

Mohr, D. C., Cuijpers, P., and Lehman, K. (2011). Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. *J. Med. Internet Res.* 13:e30. doi: 10.2196/jmir.1602

Morgan, V., Waterreus, A., Jablensky, A., MacKinnon, A., McGrath, J., Carr, V., et al. (2011). People living with psychotic illness in 2010: the second Australian national survey of psychosis. *Aust. N.Z. J. Psychiatry* 46, 735–752. doi: 10.1177/0004867412449877

O'Keeffe, G. S., and Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Am. Acad. Pediatr.* 127, 800–804. doi: 10.1542/peds.2011-0054

Oppy, G., and Dowe, D. (2016). "The turing test," in *Standford Encyclopedia of Philosophy*, ed N. Z. Edward. Available online at: http://plato.stanford.edu/archives/spr2016/entries/turing-test/

Pew Research Center. (2014). *Pew Internet and American Life Project*. Washington, DC. Available online at: http://assets.pewresearch.org/wp-content/uploads/sites/14/2014/02/12-internet-users-in-2014.jpg

Rice, S., Gleeson, J., Davey, C., Hetrick, S., Parker, A. G., Lederman, R., et al. (2016). Moderated online social therapy for depression relapse prevention in young people: pilot study of a "next generation" online intervention. *Early Interv. Psychiatry*. doi: 10.1111/eip.12354. [Epub ahead of print].

Rice, S. M., Goodall, J., Hetrick, S. E., Parker, A. G., Gilbertson, T., Amminger, G. P., et al. (2014). Online and social networking interventions for the treatment of depression in young people: a systematic review. *J. Med. Internet Res.* 16, 1–16. doi: 10.2196/jmir.3304

Rickwood, D., Telford, N., Mazzer, K., Parker, A., Tanti, C., and McGorry, P. (2015). The services provided to young people by headspace centres in Australia. *Med. J. Aust.* 202, 533–537. doi: 10.5694/mja14.01695

Rickwood, D. J., Telford, N. R., Parker, A. G., Tanti, C. J., and McGorry, P. D. (2014). Headspace - Australia's innovation in youth mental health: who are the clients and why are they presenting? *Med. J. Aust.* 200, 108–111. doi: 10.5694/mja13.11235

Saeed, H. (2016). Chatbot's life. Best place to learn about bots. *Developing a Chatbot? Learn the Difference Between AI, Machine Learning, and NLP.* (Accessed March 20, 2017). Available online at: https://chatbotslife.com/developing-a-chatbot-learn-the-difference-between-ai-machine-learning-and-nlp-40a3f745aec4#.uooug0b3d

Seligman, M. E. P., Rashid, T., and Parks, A. C. (2006). Positive psychotherapy. *Am. Psychol.* 61, 774–788. doi: 10.1037/0003-066X.61.8.774

Shah, H. (2006). A.L.I.C.E.: an ACE in digitaland. *triple C* 4, 284–292.

Skowron, M., Rank, S., Garcia, D., and Hołyst, J. A. (2017). "Zooming in: studying collective emotions with interactive affective systems," in *Cyberemotions, Understanding Complex Systems,* ed J. A. Hołyst (Springer International Publishing Switzerland), 279–304.

Smola, A., and Vishwanathan, S. V. N. (2008). *Introduction to Machine Learning*. Cambridge: University Press.

Talbot, D. (2012). Wrist sensor tells you how stressed out you are. *MIT Technol. Rev.* Available online at: https://www.technologyreview.com/s/508716/wrist-sensor-tells-you-how-stressed-out-you-are/

Tantam, D. (2006). The machine as psychotherapist: impersonal communication with a machine. *Adv. Psychiatr. Treat.* 12, 416–426. doi: 10.1192/apt.12.6.416

Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54. doi: 10.1177/0261927X09351676

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 49, 433–460. doi: 10.1093/mind/LIX.236.433

Van Gemert-Pijnen, J. E. W. C., Kelders, S. M., and Bohlmeijer, E. T. (2014). Understanding the usage of content in a mental health intervention for depression: an analysis of log data. *J. Med. Internet Res.* 16, 1–16. doi: 10.2196/jmir.2991

Verber, A. (2016). *Chat-Bots vs. Google Search*. (Accessed August 19, 2016). Available online at: https://medium.com/swlh/chat-bots-vs-google-search-b44d11a53f14#.mzypias25

Wadley, G., Lederman, R., Gleeson, J., and Alvarez-Jimenez, M. (2013). "Participatory design of an online therapy for youth mental health," in *OZCHI'13 Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (Adelaide, SA). doi: 10.1145/2541016.2541030

Weisband, S., and Kiesler, S. (1996). "Self-disclosure on computer forms: meta-analysis and implications," in *CHI 96 - Electronic Proceedings*. Available online at: http://www.sigchi.org/chi96/proceedings/papers/Weisband/sw_txt.htm doi: 10.1145/238386.238387

Weizenbaum, J. (1966). ELIZA a computer program for the study of natural language communication between man and machine. *ACM Commun* 9, 36–45. doi: 10.1145/365153.365168

Ratings of Perceived Exertion and Self-reported Mood State in Response to High Intensity Interval Training. A Crossover Study on the Effect of Chronotype

Jacopo A. Vitale[1], Antonio La Torre[2], Roberto Baldassarre[3], Maria F. Piacentini[3] and Matteo Bonato[2]*

[1] Laboratory of Biological Structures Mechanics, Istituto Ortopedico Galeazzi (IRCCS), Milan, Italy, [2] Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milan, Italy, [3] Functional Evaluation and Analysis of Sport Performance, Department of Movement, Human and Health Sciences, Foro Italico University of Rome, Rome, Italy

The aim of this study was to investigate the influence of chronotype on mood state and ratings of perceived exertion (RPE) before and in response to acute high intensity interval exercise (HIIE) performed at different times of the day. Based on the morningness–eveningness questionnaire, 12 morning-types (M-types; $N = 12$; age 21 ± 2 years; height 179 ± 5 cm; body mass 74 ± 12 kg) and 11 evening-types (E-types; $N = 11$; age 21 ± 2 years; height 181 ± 11 cm; body mass 76 ± 11 kg) were enrolled in a randomized crossover study. All subjects underwent measurements of Profile of Mood States (POMS), before (PRE), after 12 (POST12) and 24 h (POST24) the completion of both morning (08.00 am) and evening (08.00 p.m.) training. Additionally, Global Mood Disturbance and Energy Index (EI) were calculated. RPE was obtained PRE and 30 min POST HIIE. Two-way ANOVA with Tukey's multiple comparisons test of POMS parameters during morning training showed significant differences in fatigue, vigor and EI at PRE and POST24 between M-types and E-types. In addition, significant chronotype differences were found only in POST12 after the evening HIIE for fatigue, vigor and EI. For what concerns Borg perceived exertion, comparing morning versus evening values in PRE condition, a higher RPE was observed in relation to evening training for M-types ($P = 0.0107$) while E-types showed higher RPE values in the morning ($P = 0.008$). Finally, intragroup differences showed that E-types had a higher RPE respect to M-types before ($P = 0.002$) and after 30 min ($P = 0.042$) the morning session of HIIE. No significant changes during the evening training session were found. In conclusion, chronotype seems to significantly influence fatigue values, perceived exertions and vigor in relation to HIIE performed at different times of the day. Specifically, E-types will meet more of a burden when undertaking a physical task early in the day. Practical results suggest that performing a HIIE at those times of day that do not correspond to subjects' circadian preference can lead to increased mood disturbances and perceived exertion. Therefore, an athlete's chronotype should be taken into account when scheduling HIIE.

INTRODUCTION

Chronotype, also defined circadian typology, represents the expression of an individual's circadian rhythmicity. Three different categories of chronotype can be defined: evening-types (E-types), morning-types (M-types), and neither-types (N-types). Typically, the chronotype is determined with the use of self-assessment questionnaire and the most used is the Morningness–Eveningness Questionnaire (MEQ) by Horne and Ostberg (1976). The existing evidence suggests that chronotype widely affects our biological, behavioral and psychological functions (Adan et al., 2012; Vitale et al., 2017a). M-types, for instance, show an early peak along the day of body temperature (Baehr et al., 2000), serum cortisol (Bailey and Heitkemper, 2001) or blood melatonin circadian rhythm (Mongrain et al., 2004), and they usually perform best in the morning (Rossi et al., 2015). On the contrary, E-types show delayed circadian acrophases compared to M-types in a range of 1–3 h and they have better performances in the evening (Montaruli et al., 2017; Roveda et al., 2017).

The chronotype also influences the individual's behavioral circadian parameters. A strong association between circadian typology and sleep–wake behavior has been observed (Vitale et al., 2017b): E-types show difficulty in initiating sleep and they usually wake up and go to bed later (Taillard et al., 2004), whereas M-types have early bedtimes, wake up times and show higher objective and subjective sleep quality (Vitale et al., 2015). Furthermore, it is crucial to emphasize that individual differences, meant as the predisposition toward morningness or eveningness, also affect the psychological functioning and the personality (Cavallera and Giudici, 2008). Moods range on a continuum from pleasurable to unpleasant feeling states along the day and it has been demonstrated that M-type women have lower levels of anxiety than E-types and N-types (Muro et al., 2009) whereas evening-oriented people presented higher values of work-related fatigue (Martin et al., 2012). In addition, E-types generally have higher scores in extraversion than M-types (Langford and Glendon, 2002) and a positive relationship between agreeableness/conscientiousness and morningness was observed (Randler, 2008).

Studies on circadian rhythms of perceived exertion, anxiety and mood states in response to physical activity are extremely limited and unclear and it is commonly claimed that mood changes in response to exercise are not influenced by time of day (Trine and Morgan, 1995). It was observed, in male adults, that the state of anxiety, vigor and anger seem to be reduced post-exercise when compared with pre-exercise levels, regardless of the time of day (O'Connor and Davis, 1992; Koltyn et al., 1998), and this result suggests that physical activity can have large effects on mood (Lattari et al., 2014). Noteworthy, it should be stressed that no previous study took into account the subjects' circadian typology in the study of the psychological responses to exercise at different times of day.

Recently, an association between chronotype, ratings of perceived exertion (RPE), fatigue scores and mood states has been observed (Vitale et al., 2013; Kunorozva et al., 2014). It seems that M-types have more of an advantage in the morning because they are less fatigued in the first hours of the day compared with N- and E-types (Rossi et al., 2015). Rae et al. (2015) reported a significant influence of chronotype on both fatigue and vigor in relation to a maximum-intensity physical task. The authors compared 200-m time-trial swimming performance, RPE and mood state at 06:30 a.m. and 6:30 p.m. in 26 swimmers, classified as 15 M-types and 11 N-types. The Profile of Mood States (POMS) questionnaire, which is a reliable and valid measure of mood in sport settings too (Terry and Lane, 2000), was used to assess the subjects' affective and mental state (McNair et al., 1971). M-type swimmers reported lower fatigue and higher vigor scores before the 06:30 time trial compared with the 6:30 p.m. and, in addition, they showed lower global mood disturbance (GMD) compared with N-types, irrespective of the time of day.

A recent systematic review deeply examined the effect of chronotype on both the results of, and the psychophysiological responses to, physical activity. The authors concluded that M-types have, in general, both better athletic performances and lower fatigue scores in the morning than N-types and E-types, especially during submaximal and self-paced physical tasks (Vitale and Weydahl, 2017). Nonetheless, few data are available about the chronotype effect on high intensity interval exercises (HIIE). Vitale et al. (2017b) examined, for the first time, actigraphy-based sleep parameters in different chronotypes in relation to two acute sessions of HIIE performed at different times of the day. It was observed that sleep quality was poorer for M-types than E-type soccer players only after the evening training session. In addition, Bonato et al. (2017a,b) highlighted that E-types had a higher peak of salivary cortisol, higher heart rate and higher vagal indices with a significant lower parasympathetic tone respect to M-types when performing a HIIE early in the morning.

It is extremely important to understand the relationship between mood states and physical performance since one variable can influence the other and vice versa. To the best of our

knowledge, no previous study examined the chronotype effect on mood in response to HIIE.

Therefore, in the present work, we aimed to study the influence of chronotype on mood state and RPE both before and in response to acute HIIE performed at 08:00 in the morning and at 08:00 in the evening. We hypothesize that M-types have higher vigor scores and lower values of fatigue, depression, anger, tension, mood disturbance and RPE before and after morning exercise than E-types and, on the contrary, that E-types have better psycho-biological responses to HIIE after the evening session.

MATERIALS AND METHODS

Subjects

Sport science student of the School of Sport Science of the Università degli Studi di Milano, Milan, Italy were recruited for the present study during the academic year 2015–2016 ($N = 547$; 389 males and 158 females). Inclusion criteria for subject's participation to the study were: age ≥ 18 years; male; being physically active; at least 6 h of training a week and with a morning or evening chronotype scores (see "assessment of circadian typology"). Exclusion criteria were smoking, use of medications and any other medical condition contraindicating physical exercise. Thirty-seven healthy collegiate male students were therefore deemed eligible. Nevertheless, only 24 subjects (12 M-Types and 12 E-Types) agreed to voluntarily participate in the study. Before entering the study, the participants were fully informed about the study aims and procedures, and written informed consent was obtained before testing. The study protocol was approved by the Institutional Ethics Review Committee (approved on 12/10/15, prot. N. 52/15) in accordance with current national and international laws and regulations governing the use of human subjects (Declaration of Helsinki II). This trial was registered at Australian New Zealand Clinical Trials Registry (ACTRN12617000432314). After a baseline anthropometric evaluation, subjects underwent a yo-yo intermittent recovery test level 1 (Bangsbo, 1994) and then they were randomly assigned in a 1:1 ratio according to their chronotype to either morning training (Group A: $N = 12$; age 23 ± 3 years; height 175 ± 7 cm; body mass 73 ± 10 kg, weekly training volume 8 ± 2 h) that started performing the HIIT protocol at 08.00 a.m. or evening training (Group B: $N = 12$; age 21 ± 3 years; height 176 ± 5 cm; body mass 75 ± 11 kg, weekly training volume 8 ± 3 h) that started performing the HIIT protocol at 08.00 p.m. Both groups were blinded about the aim of the study.

Study Design

This was a randomized crossover study which was carried out in spring, between March and April 2016, over a period of 4 weeks. The experimental design consisted of the following: Group A performed the morning training session at 08:00 a.m. while Group B performed the evening training session at 08:00 p.m.; after a recovery period of 7 days during which subjects maintained their habitual lifestyle without performing physical

training, Group A trained in the evening while Group B trained in the morning. The study flowchart is illustrated in **Figure 1**.

One participant in Group B was excluded from analysis because he did not perform the second training session. In each test session measurement of RPE, psychophysiological recovery, and POMS were performed. RPE was performed before and after 30 min HIIT. Measurements regarding psychophysiological recovery were performed after 12 and 24 h after HIIT. All subjects were previously familiarized with all testing procedures.

Procedures

Assessment of Subject's Circadian Typology

Participants' circadian typology was assessed by the Horne-Ostberg MEQ (Horne and Ostberg, 1976). According to the MEQ-score, participants were categorized as Morning-type (scoring ≥ 59); Evening-type (scoring ≤ 41) and Neither-type, scoring (42–58). Individual chronotype scores and categories were communicated to the participants only after the completion of the experimentation.

Anthropometric Assessment

Anthropometric variables included body mass and stature. Stature was measured with a stadiometer and body mass with a portable scale to the nearest 0.5 cm and 0.1 kg, respectively (Seca 217, Vogel & Halke, Hamburg Germany). Body mass index (BMI) was calculated using the standard formula.

Yo-Yo Intermittent Recovery Test Level 1

The Yo-Yo intermittent recovery test Level 1 (Yo-Yo IR1) consisted of repeated 2×20-m runs back and forth between the starting, turning, and finishing line at a progressively increasing speed controlled by audio bleeps from a tape recorder (Bangsbo, 1994). The test protocol started with 4 running bouts at 10–13 km . h^{-1} (0–160 m) and another 7 runs at 13.5–14.0 km . h^{-1}, and thereafter speed was increased with a stepwise 0.5 km . h^{-1} speed increment every 8 running bouts (i.e., after 760, 1080, 1400, 1720-m etc.) until exhaustion. Between each running bout, the subjects had a 10 s active rest period, consisting of 2×5-m of jogging. When the subjects failed twice to reach the finishing line in time, the distance covered was recorded and represented the test result. Tests were performed on the field of an outdoor 400-m track, marked by cones, (1.22 m width and a 20-m length). Another cone placed 5 m behind the finished line marked the running distance during the active recovery period. All tests were conducted from 11:00 a.m. to 02:00 p.m. which is considered an intermediate time of day, and in dry, windless weather conditions with a temperature of about 15–20°C. Before the test, all subjects performed a standardized warm-up at the speed of the first four running bouts of the test. The total duration of the test was 6–20 min. All subjects were previously familiarized with the test, by at least one pre-test. Heart rate was recorded beat-to-beat using a Polar RS800 heart rate monitor (Polar, Kempele, Finland) in order to measure directly the HR$_{peak}$ reached during the test.

High-Intensity Interval Exercise Protocol

The HIIE protocol consisted of 4 bouts of 4 min at 90–95% HR$_{peak}$ with 3 min of active recovery at 50–60% HR$_{peak}$

FIGURE 1 | Flowchart for the study design.
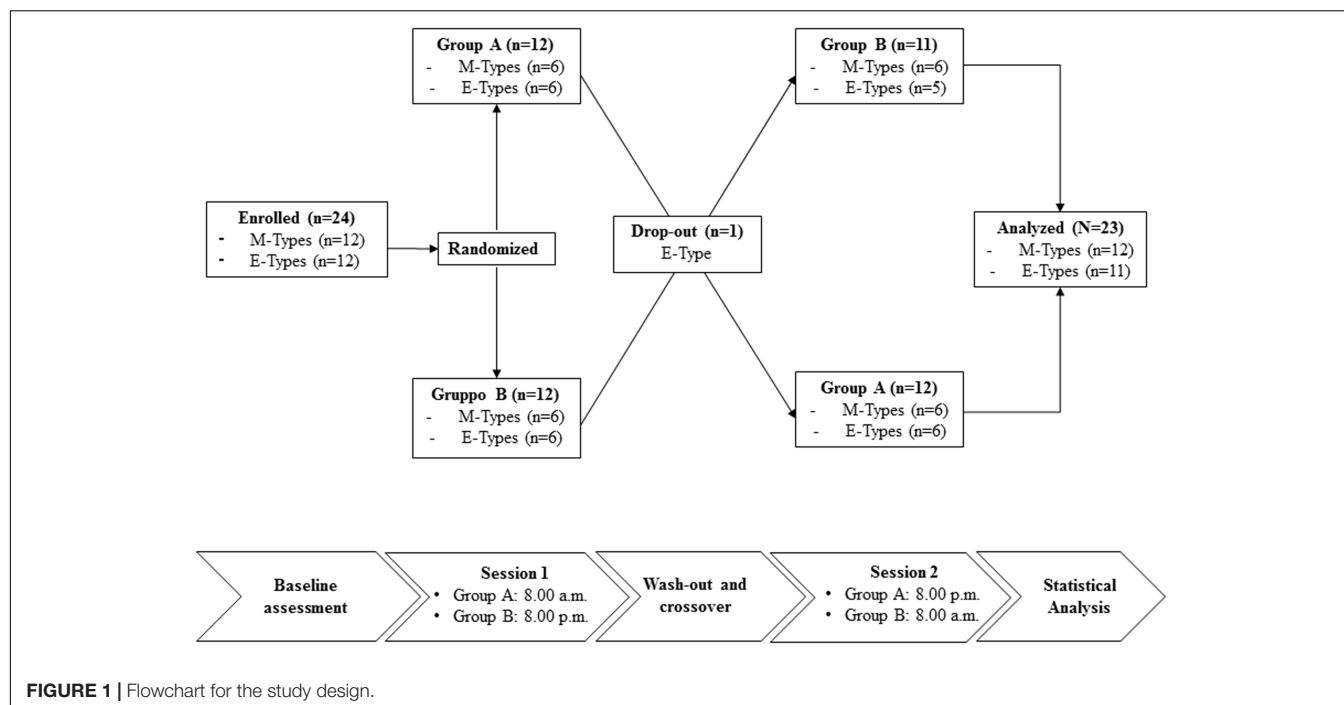
(Helgerud et al., 2001). The calculation of the training percentages was carried out using the HR_{peak} achieved during the Yo-Yo IR1. The training intervention started with a standardized 10-min warm-up and ended with a 3-min cool-down period at a self-selected intensity. Before, during, and after the test HR was recorded beat-to-beat using a Polar RS800 heart rate monitor (Polar, Kempele, Finland). All training sessions were conducted on an outdoor 400 m track in dry, windless weather conditions with a temperature of about 15–20°C. All subjects completed training sessions without complications. The high-intensity endurance interval training protocol was generally well tolerated and subjects did not report dizziness, light-headiness or nausea, symptoms that occasionally occur during this type of training.

Rating of Perceived Exertion
The Borg CR-10 category-ratio scale was selected to rate the perceived intensity of exertion (Borg, 1998). A verbal-anchored scale was shown to the subjects before (PRE), and after 30-min (POST) completing HIIE. Each subject was familiarized with the Borg CR-10 scale, including anchoring procedures.

Psychological Profile Monitoring
To evaluate the POMS a validated 32-item Italian version (Piacentini et al., 2009) questionnaire reflecting the individuals, mood on five primary dimensions (i.e., depression, fatigue, vigor, tension, and anger) was administered. Athletes were required to describe their mood (depression, fatigue, vigor, tension, and anger) using a 5-point scale (i.e., not at all = 0; somewhat = 1; moderately so = 2; very much so = 3; very very much so = 4). The questionnaires were completed individually, PRE, POST 12 and POST 24 h HIIE. An investigator was present to provide assistance if required. The POMS yields measures of

depression, fatigue, vigor, anger and tension. Data were analyzed separately for each specific dimension. Additionally, GMD was calculated by subtracting the vigor score from the sum of the scores of the four remaining subscales. To prevent a negative score, a constant of 100 was added to the global score, in accordance with Morgan et al. (1987). Given that vigor and fatigue are scores that show the greatest changes in response to training (Meeusen et al., 2013), the "energy index" (vigor-fatigue) was used to monitor these changes (Raglin et al., 1991).

Statistical Analysis
Descriptive statistics (mean ± SD) for the outcome measures were calculated. The normality of the distribution of the anthropometric (weight, height, and BMI), background (age, training hours per week, and years of practice), and Yo-Yo IR1 (total distance and HR_{peak}) variables were checked using graphical methods and the D'Agostino Pearson test. Since all variables were normally distributed, differences between Group A and Group B were checked using an unpaired Student's t-test. Parametric statistical tests were also applied to compare the POMS parameters and Borg perceived exertions, when the hypothesis of Gaussian distribution could be assumed. Specifically, intra- and inter-group differences between M-types and E-types were checked using 2-way analysis of variance with Tukey's multiple comparisons test. A paired t-test was used to compare Borg perceived exertions between morning and evening training in both PRE and POST conditions for M-types and E-types. The level of statistical significance was set at $P < 0.05$. Statistical analysis was performed using GraphPad Prism version 6.00 for Mac OSX (GraphPad Software, San Diego, CA, United States). Standardized changes in the mean values were used to

TABLE 1 | Subjects' characteristics at baseline.

Parameter	Group A	Group B	p	ES
Age (years)	23 ± 3	21 ± 3	n.s.	0.6
Height (cm)	175 ± 7	176 ± 5	n.s.	<0.2
Body Mass (kg)	73 ± 10	75 ± 11	n.s.	0.2
BMI (kg . m^{-2})	23 ± 2	23 ± 3	n.s.	<0.2
MEQ-score (points)	45 ± 16	43 ± 16	n.s.	0.2
Weekly training volume (hours . week^{-1})	8 ± 2	8 ± 3	n.s.	<0.2

Values are expressed as mean ± SD. MEQ, morningness eveningness questionnaire; ES: Effect Size.

assess magnitude of effects (Effect Size, ES). Values < 0.2, <0.6, <1.2 and >2.0 were interpreted as trivial, small, moderate, large and very large, respectively (Batterham and Hopkins, 2006).

RESULTS

Of the total of 547 subjects (71.1% males and 28.9% females), 345 were N-types (63.1%, 250 males and 95 females), 157 E-types (28.7%, 117 males and 40 females), and 45 M-types (8.2%, 22 males and 23 females). The mean MEQ score, for the whole group, was 47.2 ± 11.5 with a median of 47, 1st quartile of 38.75 and 3rd quartile of 56. The subgroup of 24 subjects was composed by 12 M-types (all moderate M-types) and 12 E-types (3 extreme E-types and 9 moderate E-types). The mean MEQ scores for the subsamples of M-types and E-types were, respectively, 31 ± 3 and 63 ± 3.

Table 1 reports the pre-HIIE parameters of the 23 subjects divided in Group A ($N = 12$) and Group B ($N = 11$), respectively. Un-paired *t*-test showed that groups were equally matched, showing no significant differences in age, height, body mass, BMI, MEQ-Score and weekly training volume.

Chronotype Effect: M-Types vs. E-Types
Morning HIIE
Table 2 shows the two-way ANOVA with Tukey's multiple comparisons test with associated *P*-values of POMS parameters during morning HIIE. A significant interaction at PRE and POST24 for fatigue, vigor and EI, with differences between M-types and E-types was found. No significant interactions were found for depression, tension, anger and GMD.

Evening HIIE
Table 3 shows the two-way ANOVA with Tukey's multiple comparisons test with associated *P*-values of POMS parameters during evening HIIE. A significant interaction at POST12 for fatigue, vigor and EI, with differences between M-types and E-types was found. No significant interactions were found for depression, tension, anger and GMD.

Figure 2 shows the POMS Iceberg Profile in relation to morning and evening training, whereas Figure 3 refers to the EI values of both training sessions. The significant differences

reported in **Figures 2**, **3** are based on 2-way analysis of variances, showed in **Tables 2**, **3**.

Rating of Perceived Exertion
Comparing morning versus evening PRE HIIE values, a higher RPE was observed in the evening for M-types (0.3 ± 0.3 vs. 1.7 ± 1.1, $P = 0.0107$, ES = 1.2) while, conversely, E-types reported higher RPE values before the start of the morning training session (2.4 ± 0.9 vs. 0.7 ± 0.9, $P = 0.008$, ES = 1.8) (**Figure 4**). No significant differences in POST condition were detected. Furthermore, as expected, RPE increased significantly 30 min post HIIE for both morning (M-types: 0.3 ± 0.4 vs. 4.5 ± 2.1, $P < 0.0001$, ES > 2.0; E-Types: 2.4 ± 0.8 vs. 6.0 ± 1.1, $P < 0.0001$, ES > 2.0) and evening (M-types: 1.2 ± 1.1 vs. 4.2 ± 2.7, $P = 0.001$, ES > 2.0; E-Types: 0.7 ± 0.9 vs. 4.4 ± 2.3, $P = 0.001$; ES > 2.0) training bouts. In conclusion, intragroup differences showed that during morning HIIE E-types had a higher RPE respect to M-types before (2.4 ± 0.9 vs. 0.3 ± 0.3, $P = 0.002$, ES > 2.0) and after 30 min (6.0 ± 1.1 vs. 4.5 ± 2.1, $P = 0.042$, ES = 1.6) HIIE. No significant changes during evening HIIE were found.

DISCUSSION

The main finding of the present study is that both RPE and mood states responses to an acute session of HIIE performed at different times of the day, are influenced by the subject's chronotype. Specifically, E-types are more fatigued, show less vigor and perceive more exertion in relation to a morning session of HIIE compared both to the evening training session and with M-types. On the contrary, mood state and scores of perceived exertions did not vary in relation to the evening session of HIIE, only M-types reported higher RPE at 08:00 p.m. compared to their morning values. It seems that performing a HIIE in the first hours of the day could generate mood disturbances and negatively influence the psychophysiological responses to physical activity for E-types but not for M-types.

It is known that morningness scores tend to increase with age (Merikanto et al., 2012) and that males are significantly more evening-oriented than females (Adan et al., 2012). Since we recruited young college students with a large predominance of males (71.1%), we observed, as expected, a larger number of E-types (28.7%) than M-types (8.2%). These results are totally in line with previous studies that investigated the chronotype distribution among young students (Adan et al., 2012; Vitale et al., 2015). One of the strengths of this work is the clear homogeneity of the sample. The participants recruited were 23 healthy and physically active male college students, categorized in 12 M-types and 11 E-types, and they were totally comparable for age, height, weight, BMI and weekly training volume, both when randomly grouped in group A and group B (**Table 1**) and also when divided for chronotype category.

Despite the literature concerning the chronotype effect on athletic performance and the psychophysiological

TABLE 2 | Results of the 2-way analysis of variance with Tukey's multiple comparisons test of the seven POMS parameters during morning HIIE for M-types and E-types at PRE, POST12 and POST24.

POMS parameters	PRE	POST 12	POST 24	Interaction	Effect of time	Chronotype effect	Intergroup differences
Depression	M: 9.1 ± 1.4 E: 10.1 ± 2.2	M: 9.4 ± 2.6 E: 8.2 ± 0.4	M: 9.0 ± 2.7 E: 8.4 ± 0.9	n.s.	n.s.	n.s.	PRE: P = n.s., ES = 0.7 POST 12: P = n.s., 0.5 POST 24: P = n.s., 0.2
Fatigue	M: 7.6 ± 2.6 E: 13.6 ± 5.6	M: 10.1 ± 3.4 E: 10.2 ± 5.6	M: 8.1 ± 2.8 E: 13.4 ± 4.8	0.028	n.s.	0.002	PRE: P = 0.024, ES > 2.0 POST 12: P = n.s., ES < 0.2 POST 24: P = 0.045, ES = 1.9
Vigor	M: 17.4 ± 4.9 E: 12.0 ± 2.1	M: 13.8 ± 4.6 E: 13.4 ± 2.5	M: 15.7 ± 3.8 E: 10.2 ± 3.5	0.048	n.s.	0.0002	PRE: P = 0.023, ES = 1.1 POST 12: P = n.s., ES < 0.2 POST 24: P = 0.011, ES = 1.4
Tension	M: 8.5 ± 2.7 E: 8.2 ± 2.8	M: 7.6 ± 3.9 E: 6.3 ± 0.7	M: 7.5 ± 4.2 E: 6.8 ± 1.6	n.s.	n.s.	n.s.	PRE: P = n.s., ES < 0.2 POST 12: P = n.s., ES = 0.4 POST 24: P = n.s., ES = 0.6
Anger	M: 8.8 ± 2.9 E: 9.4 ± 2.7	M: 9.8 ± 2.9 E: 7.3 ± 0.7	M: 9.5 ± 5.1 E: 9.0 ± 1.9	n.s.	n.s.	n.s.	PRE: P = n.s., ES = 0.2 POST 12: P = n.s., ES = 0.8 POST 24: P = n.s., ES < 0.2
Global mood disturbance	M: 33.1 ± 4.7 E: 32.7 ± 5.7	M: 32.9 ± 4.2 E: 29.1 ± 6.1	M: 32.4 ± 7.9 E: 30.2 ± 5.1	n.s.	n.s.	n.s.	PRE: P = n.s., ES < 0.2 POST 12: P = n.s. ES = 0.9 POST 24: P = n.s., ES = 0.3
Energy Index	M: 9.7 ± 6.5 E: −1.2 ± 7.7	M: 2.7 ± 6.8 E: 3.2 ± 6.4	M: 8.1 ± 5.1 E: −2.1 ± 8.7	0.015	n.s.	0.0002	PRE: P = 0.006, ES = 1.7 POST 12: P = n.s., ES < 0.2 POST 24: P = 0.012, ES = 2.0

M, M-types; E, E-types; n.s., not significant; ES, Effect Size.

TABLE 3 | Results of the 2-way analysis of variance with Tukey's multiple comparisons test of the seven POMS parameters during evening HIIE for M-types and E-types at PRE, POST12 and POST24.

POMS parameters	PRE	POST 12	POST 24	Interaction	Effect of time	Chronotype effect	Intergroup differences
Depression	M: 11.0 ± 3.7 E: 8.8 ± 2.4	M: 9.4 ± 2.9 E: 8.9 ± 1.9	M: 9.3 ± 2.4 E: 8.1 ± 0.3	n.s.	n.s.	n.s.	PRE: P = n.s., ES = 0.5 POST 12: P = n.s., ES = 0.2 POST 24: P = n.s., ES = 0.5
Fatigue	M: 8.2 ± 1.6 E: 9.7 ± 4.4	M: 8.4 ± 1.8 E: 12.9 ± 4.8	M: 10.1 ± 3.0 E: 8.1 ± 2.1	0.005	n.s.	n.s.	PRE: P = n.s., ES = 0.9 POST 12: P = 0.019, ES > 2.0 POST 24: P = n.s., ES = 0.6
Vigor	M: 15.8 ± 3.8 E: 15.0 ± 2.8	M: 16.1 ± 3.4 E: 10.8 ± 4.4	M: 12.9 ± 4.4 E: 14.6 ± 2.9	0.009	n.s.	n.s.	PRE: P = n.s., ES = 0.2 POST 12: P = 0.019, ES = 1.5 POST 24: P = n.s., ES = 0.4
Tension	M: 8.5 ± 3.9 E: 7.4 ± 2.5	M: 7.2 ± 1.7 E: 7.1 ± 2.2	M: 8.3 ± 3.4 E: 6.2 ± 0.4	n.s.	n.s.	n.s.	PRE: P = n.s., ES = 0.3 POST 12: P = n.s., ES < 0.2 POST 24: P = n.s., ES = 0.6
Anger	M: 8.3 ± 1.4 E: 8.2 ± 2.9	M: 8.0 ± 1.3 E: 10.0 ± 5.3	M: 9.0 ± 2.1 E: 7.5 ± 1.0	n.s.	n.s.	n.s.	PRE: P = n.s., ES < 0.2 POST 12: P = n.s., ES = 1.5 POST 24: P = n.s., ES = 0.7
Global mood disturbance	M: 29.8 ± 4.9 E: 31.5 ± 5.0	M: 30.3 ± 4.1 E: 31.9 ± 6.6	M: 31.0 ± 6.3 E: 28.4 ± 3.0	n.s.	n.s.	n.s.	PRE: P = n.s., ES = 0.3 POST 12: P = n.s., ES = 0.4 POST 24: P = n.s., ES = 0.4
Energy Index	M: 7.6 ± 4.7 E: 5.2 ± 6.2	M: 7.7 ± 4.2 E: −2.1 ± 8.8	M: 2.8 ± 5.9 E: 6.5 ± 4.1	0.0014	n.s.	n.s.	PRE: P = n.s., ES = 0.5 POST 12: P = 0.003, ES > 2.0 POST 24: P = n.s., ES = 0.6

M: M-types; E: E-types; n.s.: not significant; ES: Effect Size.

responses to physical activity has increased over the last years, there are still few and conflicting results. A recent systematic review highlighted that M-types have better athletic performances in the morning compared to other chronotypes (Vitale and Weydahl, 2017) but, most of all, the more evident results can be observed for RPE and fatigue scores in relation to physical activity. (Brown et al., 2008; Henst et al., 2015; Rae et al., 2015) Previous studies showed that M-types perceived less exertion when performing a moderate-intensity physical task in the morning, while E-types showed higher fatigue values in the
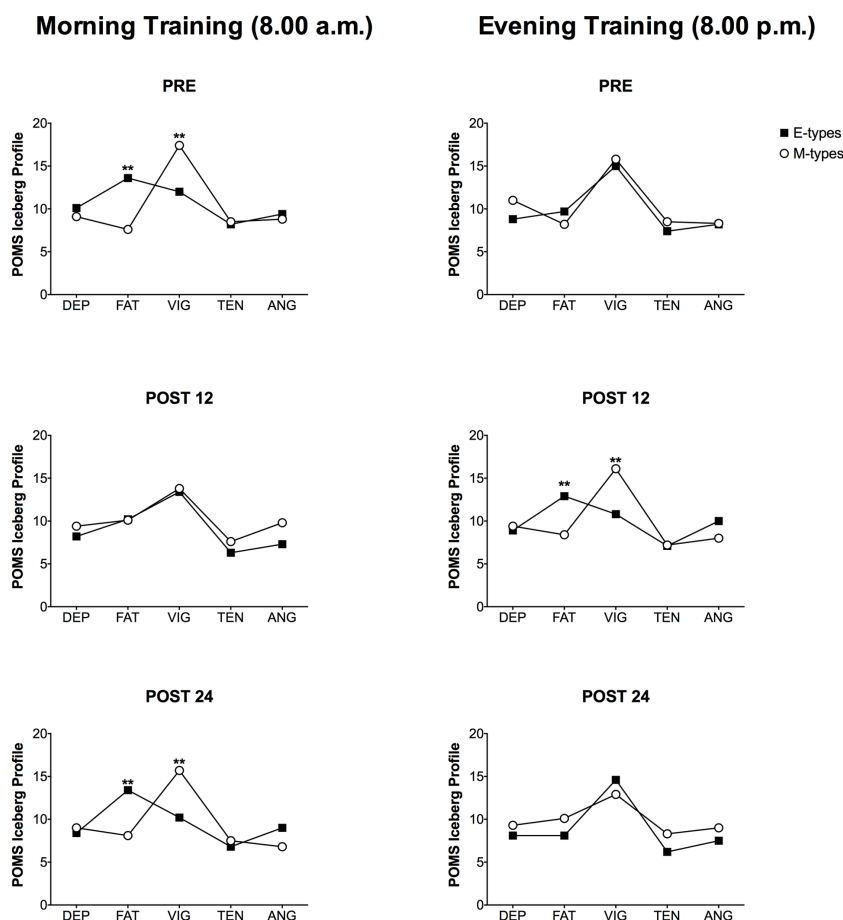
FIGURE 2 | Profile of Mood States (POMS) Iceberg Profiles of the five POMS parameters during morning and evening HIIE for M-types and E-types at PRE, POST12 and POST24. DEP, depression; FAT, fatigue; VIG, vigor; TEN, tension; ANG, anger; **$P < 0.01$. Please note that the comparison between M-types and E-types refers to the statistical analysis reported in **Tables 2**, **3**.
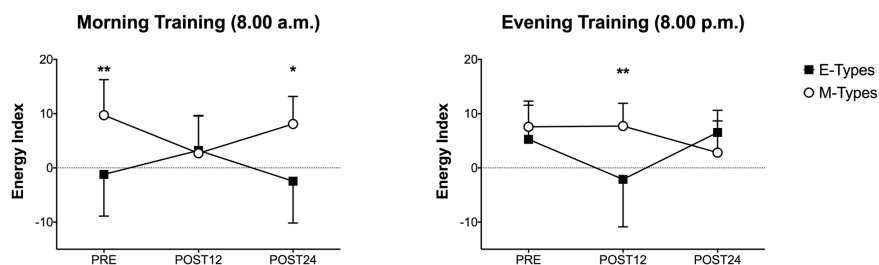
FIGURE 3 | Energy Index (EI) values during morning and evening HIIE for M-types and E-types at PRE, POST12 and POST24. *$P < 0.05$; **$P < 0.01$. Please note that the comparison between M-types and E-types refers to the statistical analysis reported in **Tables 2**, **3**.

first part of the day (Vitale et al., 2013; Kunorozva et al., 2014; Rae et al., 2015; Rossi et al., 2015).

In particular, before the start of the morning HIIE session, we observed that E-types had significantly higher RPE (2.4 ± 0.8) both than M-types (0.3 ± 0.4) and their evening values (0.7 ± 0.9). Moreover, also RPE values post HIIE performed at 08:00 remained markedly higher for E-types (6.0 ± 1.1) compared with morning subjects' scores

(4.5 ± 2.1). The only difference observed in relation to the evening HIIE session is that M-types reported higher RPE values in the PRE condition, (1.2 ± 1.1) compared to their morning values (0.3 ± 0.4). The same trend was reported by Kunorozva et al. (2014): they noted that M-type cyclists had higher RPE when cycling at 18:00 and 22:00 compared to the morning sessions. Furthermore, Rossi et al. (2015) highlighted that E-type college students had higher RPE at

FIGURE 4 | Ratings of perceived exertion (RPE) during morning and evening HIIE for M-types and E-types at PRE and POST30. *$P < 0.05$; **$P < 0.01$.
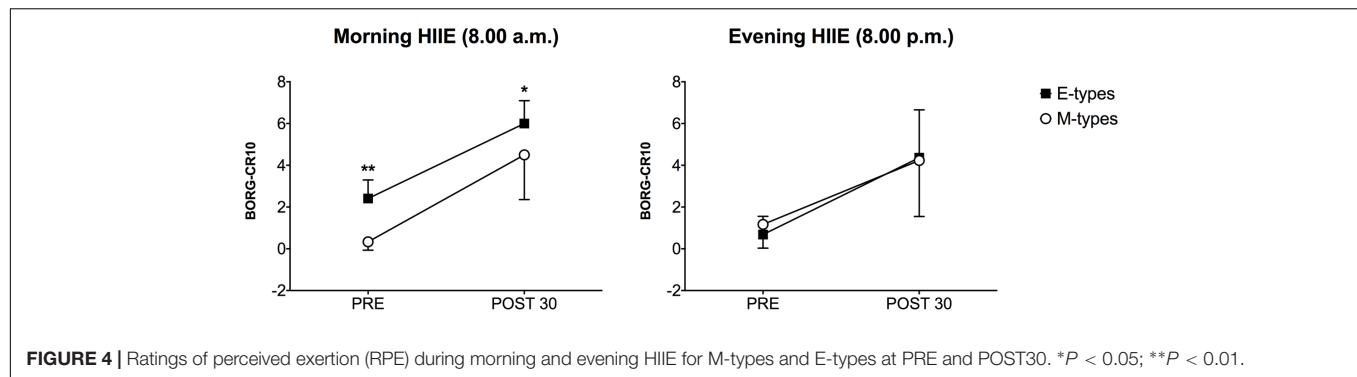
08:30 in response to a self-paced walking task compared with M-types.

To confirm this, the results of POMS questionnaire are on the same line. We observed a significant chronotype effect on vigor, fatigue and Energy Index (EI) scores: M-types had higher vigor (17.4 ± 4.9) and energy (9.7 ± 6.5) and lower fatigue (7.6 ± 2.6) than E-types (vigor: 12.0 ± 2.1; EI: -1.2 ± 7.7; fatigue: 13.6 ± 5.6) before performing the morning physical task. Nonetheless, anger, depression, tension and GMD scores did not vary according to the subjects' circadian typology. An interesting result is that the same differences were observed the morning after (POST 24), but not in the evening of the same day (POST 12) (**Figure 2** and **Table 2**). To avoid any factor that could confound the mood states the subjects were asked to lead their following day, meant as the 24 h post exercise, with their normal habits and without any kind of physical activity to not influence the POMS results. No significant chronotype effect on the POMS items was observed in PRE and POST 24 HIIE evening session (both time periods refer to evening hours). However, curiously, the morning after (POST 12), E-types reported lower vigor and EI scores and they were more fatigued compared to M-types (**Figure 2** and **Table 3**).

To the best of our knowledge, Rae et al. (2015) conducted the first and only study that evaluated the effect of chronotype on POMS items in relation with a physical task performed at different times of the day. Their results are in line with the present study: no differences were observed for GMD in accordance with the chronotype group but a significant interaction time-by-chronotype was detected for the sub-items: M-types had lower fatigue and higher vigor scores prior to the morning physical test compared to the evening session. Therefore, the lower perception of effort and greater vigor in the morning for M-types may lead them to reach better performances in the first part of the day.

All these results highlight the fact that, in general, the early hours of the day seem to represent a time that could create more disadvantage in the psychophysiological responses to HIIE, especially for E-types. Previous studies remarked this concept and reported a chronotype effect on HIIE too. Bonato et al. (2017a,b) showed that E-types had higher morning levels of salivary cortisol, heart rate and presented a significant parasympathetic withdrawal with a sympathetic predominance respect to M-types when performing a HIIE at 08:00 a.m. whereas the same

differences were not observed in the evening. Furthermore, on the contrary, Vitale et al. (2017b) reported that an evening session of high intensity interval training is more suitable for E-type collegiate soccer players: sleep quality, evaluated through actigraphy, was poorer in M-types than in evening-oriented subjects in response to the evening HIIE session.

This investigation has a number of limitations that should be discussed. First, the study population was composed by a relatively small sample and no power calculations were performed. The participants are representative only of male athletes practicing soccer while female athletes and other sports disciplines were not considered. Second, it is essential, in future studies, to control of potential confounders: it will be necessary to make appropriate decisions when selecting between field-based or laboratory-based performance tests and the differences between training sessions and official competitions should be considered.

CONCLUSION

Chronotype seems to significantly influence the psychophysiological responses to physical activity. Fatigue, perceived exertion and vigor in relation to HIIE performed at different times of the day are affected by the subjects' circadian typology. Specifically, E-types will meet more of a burden when undertaking a physical task early in the day. Practical results suggest that performing an HIIE at those times of day that do not correspond to subjects' circadian preference can lead to increased mood disturbances and perceived exertion. Therefore, an individual's chronotype should be taken into account by conditioning coaches when scheduling HIIE.

AUTHOR CONTRIBUTIONS

JV, ALT, RB, MFP, and MB, substantially contribute to the conception and in the design of the work, made the acquisition, analysis, and interpretation of data for the work. In addition, they drafted the work and revised it critically for important intellectual content. Moreover, they made the final approval of the version to be published and made the agreement for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

REFERENCES

Adan, A., Archer, S. N., Hidalgo, M. P., Di Milia, L., Natale, V., and Randler, C. (2012). Circadian typology: a comprehensive review. *Chronobiol. Int.* 29, 1153–1175. doi: 10.3109/07420528.2012.719971

Baehr, E., Revelle, W., and Eastman, C. I. (2000). Individual differences in the phase and amplitude of the human circadian temperature rhythm: with an emphasis on morningness-eveningness. *J. Sleep Res.* 9, 117–127. doi: 10.1046/j.1365-2869.2000.00196.x

Bailey, S. L., and Heitkemper, M. M. (2001). Circadian rhythmicity of cortisol and body temperature: morningness-eveningness effects. *Chronobiol. Int.* 18, 249–261. doi: 10.1081/CBI-100103189

Bangsbo, J. (1994). *Fitness Training in Football: A Scientific Approach*. Bagsværd: HO+Storm, 1–336.

Batterham, A. M., and Hopkins, W. G. (2006). Making meaningful inferences about magnitudes. *Int. J. Sports Physiol. Perform.* 9, 6–13. doi: 10.1123/ijspp.1.1.50

Bonato, M., Agnello, L., Galasso, L., Montaruli, A., Roveda, E., Merati, G., et al. (2017a). Acute modification of cardiac autonomic function of high-intensity interval training in collegiate male soccer players with different chronotype: a cross-over study. *J. Sports Sci. Med.* 16, 286–294.

Bonato, M., La Torre, A., Saresella, M., Marventano, I., Merati, G., and Vitale, J. A. (2017b). Salivary cortisol concentration after high-intensity interval exercise: time of day and the chronotype effect. *Chronobiol. Int.* doi: 10.1080/07420528.2017.1311336 [Epub ahead of print].

Borg, G. (1998). *Borg's Perceived Exertion and Pain Scales*. Champaign IL: Human Kinetics.

Brown, F. M., Neft, E. E., and LaJambe, C. M. (2008). Collegiate rowing crew performance varies by morningness–eveningness. *J. Strength Cond. Res.* 22, 1894–1900. doi: 10.1519/JSC.0b013e318187534c

Cavallera, G. M., and Giudici, S. (2008). Morningness and eveningness personality: a survey in literature from 1995 up till 2006. *Pers. Individ. Diff.* 44, 3–21. doi: 10.1016/j.paid.2007.07.009

Helgerud, J., Engen, L. C., Wisloff, U., and Hoff, J. (2001). Aerobic endurance training improves soccer performance. *Med. Sci. Sports Exerc.* 33, 1925–1931.

Henst, R. H. P., Jaspers, R. T., Roden, L. C., and Rae, D. E. (2015). A chronotype comparison of South African and Dutch marathon runners: the role of scheduled race start times and effects on performance. *Chronobiol. Int.* 32, 858–868. doi: 10.3109/07420528.2015.1048870

Horne, J. A., and Ostberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int. J. Chronobiol.* 4, 97–110.

Koltyn, K. F., Lynch, N. A., and Hill, D. W. (1998). Psychological responses to brief exhaustive cycling exercise in the morning and the evening. *Int. J. Sport Psychol.* 29, 145–156.

Kunorozva, L., Roden, L. C., and Rae, D. E. (2014). Perception of effort in morning-type cyclists is lower when exercising in the morning. *J. Sports Sci.* 32, 917–925. doi: 10.1080/02640414.2013.873139

Langford, C., and Glendon, A. (2002). Effects of neuroticism, extraversion, circadian type and age on reported driver stress. *Work Stress* 16, 316–334. doi: 10.1080/0267837021000059019

Lattari, E., Portugal, E., Moraes, H., Machado, S., Santos, T. M., and Deslandes, A. C. (2014). Acute effects of exercise on mood and EEG activity in healthy young subjects: a systematic review. *CNS Neurol. Disord. Drug Targets* 13, 972–980. doi: 10.2174/1871527313666140612125418

Martin, J. S., Hébert, M., Ledoux, E., Gaudreault, M., and Laberge, L. (2012). Relationship of chronotype to sleep, light exposure, and work-related fatigue in student workers. *Chronobiol. Int.* 29, 295–304. doi: 10.3109/07420528.2011.653656

McNair, D. M., Lorr, M., and Droppleman, L. F. (1971). *Manual for Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Services.

Meeusen, R., Duclos, M., Foster, C., Fry, A., Gleeson, M., Nieman, D., et al. (2013). European College of Sport Science; American College of Sports Medicine. Prevention, diagnosis, and treatment of the overtraining syndrome: joint consensus statement of the European College of Sport Science and the American College of Sports Medicine. *Med. Sci. Sports Exerc.* 45, 186–205. doi: 10.1249/MSS.0b013e318279a10a

Merikanto, I., Kronholm, E., Peltonen, M., Laatikainen, T., Lahti, T., and Partonen, T. (2012). Relation of chronotype to sleep complaints in the general Finnish population. *Chronobiol. Int.* 29, 311–317. doi: 10.3109/07420528.2012.655870

Mongrain, V., Lavoie, S., Selmaoui, B., Paquet, J., and Dumont, M. (2004). Phase relationship between sleep-wake cycle and underlying rhythms in morningness-eveningness. *J. Biol. Rhythms* 19, 248–257. doi: 10.1177/0748730404264365

Montaruli, A., Galasso, L., Carandente, F., Vitale, J. A., Roveda, E., and Caumo, A. (2017). If the Morning-Evening Questionnaire (MEQ) is able to predict the actigraphy-based acrophase, how does its reduced, five-item version (rMEQ) perform? *Chronobiol. Int.* 34, 443–444. doi: 10.1080/07420528.2017.1306708

Morgan, W. P., Brown, D. R., Raglin, J. S., O'Connor, P. J., and Ellickson, K. A. (1987). Psychological monitoring of overtraining and staleness. *Br. J. Sports. Med.* 21, 107–114. doi: 10.1136/bjsm.21.3.107

Muro, A., Gomà-i-Freixanet, M., and Adan, A. (2009). Morningness–eveningness, sex, and the alternative five factor model of personality. *Chronobiol. Int.* 26, 1235–1248. doi: 10.3109/07420528.2014.968282

O'Connor, P. J., and Davis, J. C. (1992). Psychobiologic responses to exercise at different times of day. *Med. Sci. Sports Exerc.* 24, 714–719.

Piacentini, M. F., Salvatori, G., Di Cesare, C., Pizzuto, F., Olivi, M., Minganti, C., et al. (2009). "Effect of zone-diet on training parameters in recreational master athletes," in *Contemporary Sport, Leisure and Ergonomics*, eds T. Reilly and G. Atkinson (London: Routledge), 227–241.

Rae, D. E., Stephenson, K. J., and Roden, L. C. (2015). Factors to consider when assessing diurnal variation in sports performance: the influence of chronotype and habitual training time-of-day. *Eur. J. Appl. Physiol.* 115, 1339–1349. doi: 10.1007/s00421-015-3109-3109

Raglin, J. S., Morgan, W. P., and O'Connor, P. J. (1991). Changes in mood states during training in female and male college swimmers. *Int. J. Sports Med.* 12, 585–589. doi: 10.1055/s-2007-1024739

Randler, C. (2008). Morningness–eveningness, sleep–wake variables and big five personality factors. *Pers. Indivd. Differ.* 45, 191–196. doi: 10.1016/j.paid.2008.03.007

Rossi, A., Formenti, D., Vitale, J. A., Calogiuri, G., and Weydahl, A. (2015). The effect of chronotype on psychophysiological responses during aerobic self-paced exercises. *Percept. Mot. Skills* 121, 840–855. doi: 10.2466/27.29.PMS.121c28x1

Roveda, E., Vitale, J. A., Montaruli, A., Galasso, L., Carandente, F., and Caumo, A. (2017). Predicting the actigraphy-based acrophase using the Morningness–Eveningness Questionnaire (MEQ) in college students of North Italy. *Chronobiol. Int.* 34, 551–5625. doi: 10.1080/07420528.2016.1276928

Taillard, J., Philip, P., Chastang, J. F., and Bioulac, B. (2004). Validation of horne and ostberg morningness-eveningness questionnaire in a middle-aged population of French workers. *J. Biol. Rhythms* 19, 76–86. doi: 10.1177/0748730403259849

Terry, P. C., and Lane, A. M. (2000). Normative values for the profile of mood states for use with athletic samples. *J. Appl Sport Psychol.* 12, 93–109. doi: 10.1080/10413200008404215

Trine, M. R., and Morgan, W. P. (1995). Influence of time of day on psychological responses to exercise. A review. *Sports Med.* 20, 328–337. doi: 10.2165/00007256-199520050-00004

Vitale, J. A., Bjoerkesett, E., Campana, A., Panizza, G., and Weydahl, A. (2017a). Chronotype and response to training during the polar night: a pilot

study. *Int. J. Circumpolar Health* 76, 1320919. doi: 10.1080/22423982.2017.132 0919

Vitale, J. A., Bonato, M., Galasso, L., La Torre, A., Merati, G., Montaruli, A., et al. (2017b). Sleep quality and high intensity interval training at two different times of day: a crossover study on the influence of the chronotype in male collegiate soccer players. *Chronobiol. Int.* 34, 260–268. doi: 10.1080/07420528. 2016.1256301

Vitale, J. A., Calogiuri, G., and Weydahl, A. (2013). Influence of chronotype on responses to a standardized, self-paced walking task in the morning vs afternoon: a pilot study. *Percept. Mot. Skills* 116, 1020–1028. doi: 10.2466/06. 19.PMS.116.3.1020-1028

Vitale, J. A., Roveda, E., Montaruli, A., Galasso, L., Weydahl, A., Caumo, A., et al. (2015). Chronotype influences activity circadian rhythm and sleep: differences in sleep quality between weekdays and weekend. *Chronobiol. Int.* 32, 405–415. doi: 10.3109/07420528.2014.986273

Vitale, J. A., and Weydahl, A. (2017). Chronotype, physical activity, and sport performance: a systematic review. *Sports Med.* doi: 10.1007/s40279-017-0741-z [Epub ahead of print].

Book Review: Networks of the Brain

Fozia Anwar, Afifa Yousafzai and Muaz A. Niazi*

COSMOSE Research Group, Health Informatics, COMSATS Institute of Information Technology, Islamabad, Pakistan

A Book Review on
Networks of the Brain

Olaf Sporns, (Cambridge, MA: MIT Press), 2010, 424 pages.

INTRODUCTION

Natural systems consisting of living systems as well as some artificial systems (Altamimi and Ramadan, 2016; Batool and Niazi, 2017) can be classified as Complex Adaptive Systems (CAS) (Holland, 2012). These systems have complexity in terms of having nonlinear interactions between numerous agents—interactions which cannot be easily summarized. CAS also have some rather peculiar properties—they adapt such that individual agents do not matter—an example being of cells which can continually get replaced with no discernable loss of overall emergent functionality. Examples of CAS include humans or animals. Such systems simply cannot be understood by dissection or only from the perspective of being made up of a wide variety of cells or elements/compounds for that matter.

CAS are all around us and even inside us. Human beings consist of numerous hierarchical and intertwined CAS both individually as well as collectively. CAS are so close and abundant on this planet that it is actually very difficult think of anything with no link to a CAS—hence we cannot see the forest for the trees. The only way to make sense of CAS seems to be to develop various types of models or simplified abstractions of complex real-world systems. Modeling takes out any unnecessary details—example being of maps (Miller and Page, 2007). Maps provide information about the real-world and yet do not cover every possible detail. Likewise, agents and networks can be used to model the real world or concepts inspired (Zedadra et al., 2017) by the real world.

Like maps, there are models which are more appropriate to modeling CAS. These include agent-based, complex network based (Mrvar and Batagelj, 2016), and multiscale models. The book which is being reviewed here applies the complex network approach to modeling the brain.

REVIEW

The goal of this book is to exemplify the importance of and usage of network science in neuroanatomy. The book gives a holistic approach to the complex network perspective including its historical roots.

The book has four portions. The first one has three chapters covering the basic details of the two domains. The second portion, consisting of chapters 4–7, focuses on the anatomical networks of cells and regions. Whereas the third portion of the book (chapters 8–11), is dedicated to network dynamics. The final portion has three chapters covering different aspects of network complexity.

Chapter 1 eloquently explains the similarities between brain networks architecture and architecture of complex systems. It also motivates some key neuroscience questions which could be effectively addressed using network models. Chapter 2 introduces the basics both in terms of terminologies as well as methodologies using in network science. It also includes as perceptive

survey of the quantitative tools and concepts which could potentially be useful for brain-related studies. Chapter 3 describes some fundamental techniques and approaches used to mine and extract brain networks from neurological data.

The second portion starts with Chapter 4 offering a network perspective to the functional anatomy of the brain including its historical provenance. An emerging set of ideas is also outlined regarding functional and structural specialization. Chapter 5 outlines modern neuroanatomical techniques useful for mining brain networks. Chapter 6 presents some so well-known key architectural principles of anatomical networks. Chapter 7 examines the functional meaning and evolutionary origins. The chapter examines the so-called "wiring minimization" hypothesis. This includes questions such as, to what extent has brain connectivity been shaped by spatial and metabolic constraints? Also, the optimal economy of the elements and connections of brain networks (in a spatial or metabolic sense) is examined besides some other questions.

The third portion of the book begins with Chapter 8 focusing on functional networks generated by spontaneous activity in neural systems. Chapter 9 attempts to draw links between brain networks and cognition. Chapter 10 outlines existing knowledge about brain network disruptions in neurological and psychiatric diseases. Chapter 11 focuses on the growth, development, and aging of brain networks. Chapter 12 makes the case for diverse and flexible neural dynamics as a prerequisite for efficient computation. Chapter 13 traces the origin of complex dynamic patterns for structural patterns of network connectivity. Finally,

chapter 14 examines the role of the body in shaping the functioning of brain networks.

CRITICAL COMMENTS

The book is generally well-written but is probably not suitable as a first book in Network Science. It is a thoroughly technical book. Easy and nonmathematical approach of this book towards the subject still make this book a good read. References used in this book are also up-to-date. An appropriate first book on networks could be the text by Newman (2010).

CONCLUSIONS

Olaf Sporns has presented a readable treatise for readers interested in complex systems as well as those interested in areas ranging from quantitative psychology to neurosciences. The idea of modeling complex neuronal structures as networks is quite new. While there are some previously published books in this domain, there was a need for a comprehensive text on this topic. Although the book does not introduce complexity-related details it is still a very interesting read and highly recommended as a ready reference.

AUTHOR CONTRIBUTIONS

MN conceived the idea of the paper. MN, FA, and AY all wrote and edited the paper. All authors read and approved the final manuscript.

REFERENCES

Altamimi, A. B., and Ramadan, R. A. (2016). Towards internet of things modeling: a gateway approach. *Complex Adapt. Syst. Model.* 4:25. doi: 10.1186/s40294-016-0038-3

Batool, K., and Niazi, M. A. (2017). Modeling the internet of things: a hybrid modeling approach using complex networks and agent-based models. *Complex Adapt. Syst. Model.* 5:4. doi: 10.1186/s40294-017-0043-1

Holland, J. H. (2012). *Signals and Boundaries: Building Blocks for Complex Adaptive Systems*. Cambridge, MA: MIT Press.

Miller, J. H., and Page, S. E. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton, NJ: Princeton University Press.

Mrvar, A., and Batagelj, V. (2016). Analysis and visualization of large networks with program package Pajek. *Complex Adap. Syst. Model.* 4:6. doi: 10.1186/s40294-016-0017-8

Newman, M. (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press.

Zedadra, O., Jouandeau, N., Seridi, H., and Fortino, G. (2017). Multi-Agent Foraging: state-of-the-art and research challenges. *Complex Adap. Syst. Model.* 5:3. doi: 10.1186/s40294-016-0041-8

Study Protocol for the Preschooler Regulation of Emotional Stress (PRES) Procedure

Livio Provenzi[1]*, Rafaela G. M. Cassiano[2], Giunia Scotto di Minico[1], Maria B. M. Linhares[2] and Rosario Montirosso[1]

[1] 0-3 Center for the at-Risk Infant, Scientific Institute IRCCS Eugenio Medea, Bosisio Parini, Italy, [2] Department of Neurosciences and Behavior, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil

Background: Emotional stress regulation (ESR) rapidly develops during the first months of age and includes different behavioral strategies which largely contribute to children's behavioral and emotional adjustment later in life. The assessment of ESR during the first years of life is critical to identify preschool children who are at developmental risk. Although ESR is generally included in larger temperament batteries [e.g., the Laboratory Temperament Assessment Battery (Lab-TAB)], there is no standardized observational procedure to specifically assess and measure ESR in preschool aged children.

Aim: Here, we describe the development of an observational procedure to assess ESR in preschool aged children [i.e., the Preschooler Regulation of Emotional Stress (PRES) Procedure] and the related coding system.

Methods: Four Lab-TAB emotional stress episodes (i.e., the Stranger, the Perfect Circle, the Missing Sticker, and the Transparent Box) have been selected. Independent coders developed a list of ESR codes resulting in two general indexes (i.e., active engagement and stress level) and five specific indexes (i.e., anger, control, fear, inhibition, sadness). Finally, specific actions have been planned to assess the validity and the coding system reliability of PRES procedure.

Ethics and Dissemination: The study has been approved by the Ethical Committee of the Scientific Institute IRCCS Eugenio Medea, Bosisio Parini (Italy). The PRES validation and reliability assessment as well as its use with healthy and at-risk populations of preschool children will be object of future scientific publications and international conference presentations.

Keywords: emotion regulation, observational methods, preschoolers, stress response, study protocol

INTRODUCTION

Background

The ability to cope with emotional stress [i.e., emotional stress regulation (ESR)] develops early during the first months of life (Thompson, 1994; DiCorcia and Tronick, 2011) and is a key component of infants and children's temperament (Rothbart et al., 2011). The early development of adequate behavioral strategies of ESR is thought to be one of the major predictive factor of emotional (Eisenberg et al., 2010; Berking and Wupperman, 2012), cognitive (Compas et al., 2014; Blankson et al., 2017), and social (Cisler and Olatunji, 2012; Pisani et al., 2013) adjustment later in life. For this reason, the availability of an observational tool to assess ESR in preschool children is of crucial importance.

The Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith et al., 1993) is a set of experimental tasks specifically developed to measure preschool children's temperament. Although it includes specific episodes which depict different ESR behavioral strategies, there is no standardized Lab-TAB procedure which allows to assess ESR in preschoolers thoroughly. In the present contribution, we present the study protocol for the development of a specific laboratory procedure to assess ESR in preschool aged children (namely, the Preschooler Regulation of Emotional Stress, PRES procedure) which is built on a comprehensive and standardized selection of specific Lab-TAB episodes which highlight behavioral strategies used to cope with emotional stress.

ESR Conceptualization
Relevance of ESR for Child Development
Human infants develop adaptive behavioral strategies to cope with emotional stressors during the first months of life and within daily interactions with their main caregiver (DiCorcia and Tronick, 2011). As mother–infant and mother–child interactions are inherently characterized by frequent ruptures and reparations (Weinberg and Tronick, 1998; Beeghly et al., 2011), emotional stress is a part of infants' interactions with their everyday human environment. Infants are thought to develop adaptive behavioral strategies to regulate emotional stress through repeated experiences of co-regulation of interactive and communicative ruptures, as they move from the need of external sources of regulation provided by the caregiver to adequate self-regulation abilities (Beeghly and Tronick, 2011; Bornstein and Manian, 2013).

Less-than-optimal strategies of ESR have been described in different subjects at-risk due to both environmental and genetic factors, including prematurity (Hsu and Jeng, 2008; Montirosso et al., 2010; Langerock et al., 2013), trauma-exposed children (Dvir et al., 2014; Szentágotai-Tătar and Miu, 2016), and infants with specific genotypes associated with stress susceptibility (e.g., the serotonin transporter polymorphism, 5-HTTLPR; Pauli-Pott et al., 2009; Montirosso et al., 2015b). Moreover, altered ESR during the early stages of development is considered to be a critical risk factor for further behavioral and affective development during adulthood. Indeed, children characterized by emotional stress dysregulation have a heightened risk of developing psychopathology during adolescence (McLaughlin et al., 2011) and adult-age (Compas et al., 2014; Huh et al., 2017).

In the light of this evidence, standardized and well-validated approaches to ESR assessment during infancy and preschool age appear to be crucial in order to better understand the consequences of early adversities on human emotional development and to adequately plan effective preventive and therapeutic interventions.

Toward a Multi-faceted and Processual Conceptualization of ESR
ESR is a multifaceted construct (Aldao et al., 2016) which includes expressions and behaviors which allow individuals to cope with emotional stressors (**Table 1**). The display of emotional distress, such as negative emotionality, has been reported in

infants (Braungart-Rieker et al., 1998), and specific emotional expressions, such as anger (Gilliom et al., 2002), fear (Buss and Kiel, 2011), and sadness (Compas et al., 2014) have been observed in preschool aged children. Gaze aversion with the source of emotional stress is another strategy used to regulate behavioral states during challenging conditions (Beebe and Steele, 2013). Furthermore, children might try to obtain external support from the adult caregiver (i.e., co-regulation) by making active attempts to obtain social engagement behaviors from the mother (Ekas et al., 2013; Provenzi et al., 2015a) as well as signals of protest (Ahnert et al., 2004; Kaitz et al., 2010). Finally, older infants and preschool aged children develop behavioral strategies aimed to achieve behavioral and emotional regulation (Planalp and Braungart-Rieker, 2015) which include attempts to obtain control over the situation as well as inhibition of the emotional reaction (Adrian et al., 2011).

Moreover, ESR is thought to be a two-step process that includes: a reactivity phase, during which the individual gathers strength face an external source of emotional stress and to activate specific behavioral outputs to respond and cope with the environmental challenge; and a recovery phase, during which the organism reaches a new homeostatic and quiet state when the emotional stress condition is over (Linden et al., 1997; Tsigos et al., 2000). In other words, adaptive ESR includes the adoption of adequate strategies to react to an external source of emotional stress (i.e., reactivity) as well as the return to baseline behavioral states when the stress is over (i.e., recovery).

Available Tools to Assess ESR in Infants and Children
ESR Assessment during Infancy
The assessment of ESR in infants has been carried out according to different observational paradigms, including frustration tasks (Buss and Goldsmith, 1998), emotion-inducing tasks (Malone et al., 1985) and structured mother–infant interactions (Feldman, 2007). Nonetheless, the Face-to-Face Still-Face (FFSF) procedure (Tronick et al., 1978) is the most used and validated procedure to obtain information on expressive and coping behaviors adopted by infants to face emotional stress during the first months of life (Mesman et al., 2009; Provenzi et al., 2016).

During the FFSF procedure, emotional stress arises from the experimental manipulation of maternal responsiveness and availability in the interaction with the infant (i.e., maternal still-face). First, after 2 min of normal face-to-face interaction (i.e., *Play* episode), mothers are instructed to interrupt any communication with the infant, to avoid physical contact and to maintain a still/poker-face while looking their infant in the eye (Tronick et al., 1978). During this FFSF episode (i.e., *Still-Face* episode) infants are expected to exhibit specific reactivity behaviors in response to emotional stress (i.e., maternal unresponsiveness) including heightened negative emotionality and avoidant behaviors as well as reduced engagement (Adamson and Frick, 2003; Montirosso et al., 2015a). After the *Still-Face*, mothers and infants resume normal face-to-face interaction (i.e., *Reunion* episode) as during the *Play* episode. The *Reunion* episode allows the observers to obtain information about

TABLE 1 | Conceptualization of ESR in infants and preschool children, with examples of the relative expressive and coping behaviors.

		Expressive behaviors	Coping behaviors
Infants	Reactivity	*e.g., Negative emotionality*	*e.g., Avoiding behaviors*
	Recovery	*e.g., Positive emotionality*	*e.g., Social engagement*
Children	Reactivity	*e.g., Negative emotions (anger, fear, sadness)*	*e.g., Inhibitory control*
	Recovery	*e.g., Positive emotionality*	*e.g., Active and social engagement*

infants' capacity to recover from emotional stress as the social engagement resumes, and the reduction of negative emotionality, despite previous research has documented that a typical carry-over effect of negative emotionality is generally observed (Yato et al., 2008; Mesman et al., 2009). As such, the FFSF procedure allows to observe infants' behavior during both the reactivity and recovery phases of ESR.

Specific coding systems have been developed and validated for the FFSF procedure [e.g., the Infant Regulatory Score System (IRSS); the Infant-Caregiver Engagement Phases (ICEP)]. These coding systems include specific indexes of ESR behavioral strategies such as infants' gaze direction, vocalizations, gestures, self-comforting behaviors, distancing behaviors, and general indexes of motor activity which can be resumed as expressive (i.e., negative and positive emotionality) as well as coping (i.e., social and object engagement) behavioral indexes. As such, the FFSF is also characterized by a comprehensive assessment of infants' ESR behavioral strategies.

Assessment of ESR in Preschool Children

Previously adopted procedures have been developed to assess different aspects of children's behavioral and emotional development. For example, the Strange Situation procedure has been applied to older infants and preschool aged children, but it has been developed to assess attachment-related behaviors. As such, despite the fact that ESR and attachment have known interconnections (Zimmer-Gembeck et al., 2015), the Strange Situation procedure might lack the adequate fine-grained sensitivity to depict different ESR behavioral strategies and it usually does not provide information on the two-step reactivity-recovery process. Other laboratory procedures, such as the frustration task (Melnick and Hinshaw, 2000), non-standardized stranger approaches (Zimmermann and Stansbury, 2004), rigged peer competition (Hughes et al., 2002), fear-inducing paradigms (Buss and Goldsmith, 1998), and frustration-inducing tasks (Stifter and Braungart, 1995; Cole et al., 2003) appear to be stand-alone tasks which only partially cover the different types of emotional stress that preschool aged children might face.

Notably, many of these individual tasks have been included in the Lab-TAB, which is a set of laboratory tasks specifically developed to measure temperament in preschool aged children. As ESR contributes to the definition of a temperamental profile of children (Rothbart et al., 2006), it is not surprising that the Lab-TAB includes specific episodes aimed at observing and assessing behavioral strategies used by preschool aged children to cope with emotional stress. Nonetheless, it should be noted that temperament represents a global account of children's behavioral trait predispositions which only partially overlap with

ESR (Fox and Calkins, 2003). Indeed, whereas temperament represents an overall behavioral tendency of children with innate biological underpinnings (Goldsmith et al., 1987; Rothbart et al., 2001; Saudino and Micalizzi, 2015), ESR appears to be much more dynamic and processual, contingent to environmental conditions (i.e., emotional stressors) and affected both by genetic predispositions (Lesch, 2011; Waider et al., 2011; Ford et al., 2014) and the quality of early caregiving environment (Bariola et al., 2011; Morris et al., 2011; Kim and Kochanska, 2017).

Limitations of the Lab-TAB to Assess Preschool Children ESR

The Lab-TAB presents a series of challenges when it comes to its application on the observation of ESR in preschool aged children. First, the Lab-TAB is not entirely specific to ESR. Although it includes emotion-eliciting episodes (Gagne et al., 2011), the available coding system is meant to provide measures of temperament (e.g., levels of activity, approach, persistence), rather than contextual regulation of emotional stress behavioral strategies. Second, the Lab-TAB is made up of more than 30 episodes (Gagne et al., 2011). Consistently, sometimes the Lab-TAB is administered in two or more sessions, and the actual duration varies according to the segmentations of the Lab-TAB procedure and to children's characteristics (e.g., age). As such, previous researchers have selected different sets of Lab-TAB episodes in their studies. For example, 20 Lab-TAB episodes have been used with 4.5-year-old children (Gagne et al., 2011), 12 episodes at with 3 year olds, nine episodes with 6 year olds (Dyson et al., 2015), and two episodes with 12-month-old infants (Zmyj et al., 2017). Third, there is no clear available rationale to guide the selection of Lab-TAB episodes, which has resulted in the proliferating of various subjective "sub-versions" of the Lab-TAB procedure. For example, effortful control has been measured using four episodes (Car Seat, Puppets, Masks, and Risk room; Kochanska and Knaack, 2003) in 14-to-22-month-old preschoolers, behavioral inhibition has been assessed using three episodes (Dinky toys, Snack delay, and Gift; Gartstein and Marmion, 2008) in 2-year-old children, positive emotionality has been coded according to five different episodes, including Puppets, Peek-a-Boo game, Pop-up bunny, Snake, Bubbles (Kochanska et al., 2007). Finally, despite the authors of the original Lab-TAB manual provided a general guide to code children's temperament, researchers that used the Lab-TAB have developed different coding systems depending on the objectives of their research projects, which has resulted in the production of indexes that are only partially comparable, e.g., interest, initiative, sociability, compliance (Dyson et al., 2015); anger, fear, shyness, approach, persistence (Gagne et al., 2011); negative reactivity

(Shanahan et al., 2008); and two global indexes of both negative and positive emotionality (Kopala-Sibley et al., 2017).

As such, although the Lab-TAB is a well-established procedure used to assess preschool aged children's temperament in a laboratory setting, a standardized protocol to guide its administration to evaluate preschoolers' behavioral strategies to cope with emotional stress is non-existent up to date.

The Present Study

In the present manuscript, we describe the PRES protocol and we provide details on (1) the theoretical and methodological reasons that guided us in choosing specific Lab-TAB stress-related episodes; (2) the procedural steps for the PRES development; (3) the operational definitions of the PRES codes and indexes; (4) the methodological steps planned to assess the coding system's validity and reliability.

MATERIALS AND EQUIPMENT

Development of the Observational Procedure

Selection of Lab-TAB Episodes: Rationale and Description

The setting for the PRES procedure is graphically schematized in **Figure 1**. An essential thesaurus of PRES-related terms is provided in **Table 2**. The PRES procedure includes four episodes (i.e., Stranger, Perfect circle, Sticker, Transparent box) which have been extracted from the original Lab-TAB. These episodes have been chosen in order to represent the different types of emotional stress which have already been targeted in previous research (Gunnar et al., 2009; Adrian et al., 2011). Moreover, they have been selected in order to guarantee an easy-to-reproduce observational setting so that the procedure can be replicated in different laboratories without the need of expensive or *ad hoc* materials. The procedural steps of the four episodes are described in detail in **Table 3**.

Emotional Stressors

The Stranger episode elicits stress due to the encounter with an unfamiliar adult that is approaching and talking to the child and with whom the child had no previous relational or interactive history. As such, this episode is meant to elicit fear-related stress.

The Perfect circle episode elicits stress due to frustration and perception of self-inefficacy within a relational framework with an adult experimenter who gives negative feedbacks about the graphical production of a circle and asks for further drawing attempts without providing guidelines on how a "real" perfect circle should be drawn.

The Sticker episode induces stress due to the mismatch between the expectation of a reward (i.e., the chosen sticker) and the lack of the desired adult response (i.e., absence of the chosen sticker). As such the Perfect circle and the Sticker episodes both elicit frustration-related stress in the child, but, whilst the first is caused by a perception of self-inadequacy (e.g., being unable to draw a perfect circle), the second is triggered by an attribution

of inadequacy to the other (e.g., being unable to maintain a promise).

Finally, the Transparent box episode elicits stress due to the simultaneous presence of a visible desirable object within a transparent container, and the impossibility to reach it. As such, this task is meant to observe the perseverance – or lack thereof – of the child's efforts to achieve a desired goal while facing an impossible task. Moreover, the emotional stress elicited in this fourth episode is related to the presence of a desirable object within sight which cannot be reached nor played with by the child.

STEPWISE PROCEDURES

Development of the Micro-analytical Coding System

Naïve Coding of Infant and Child Observation

A first set of unstructured and non-hierarchical codes was developed by two researchers (authors LP and RMC), who

FIGURE 1 | Schematic representation of the CASES setting. *The director (D) sits behind a hidden camera. The hidden camera may be behind a cloth or in a room with a unidirectional mirror next to the PRES room. The *barrier* might be a chair or a little bench, in order to avoid that the child gets to the opposite side of the table and sees the director and the camera.

TABLE 2 | Thesaurus of PRES Procedure-related terms.

Term	Description
Session	Each child participates in a session, which is a complete PRES procedure including four episodes (i.e., Stranger, Perfect circle, Stickers, Transparent box).
Episodes	The episodes are four stress-inducing experimental tasks: Stranger, Perfect circle, Stickers, Transparent box.
Segments	Each episode is subdivided in three or more segments. The number of segments varies among episodes: Stranger (five segments), Perfect circle (three segments), Stickers (six segments), Transparent box (three segments).
Phases	For each episode, the segments are grouped into three phases: baseline, reactivity, and recovery.
Epochs	The coding procedure is micro-analytical and it is done every 10-s intervals.

TABLE 3 | Description of the four PRES Procedure episodes.

#	Name	Brief description	Phases (segments)
1	Stranger approach	A male stranger gradually approaches the child physically and verbally. The stranger asks questions regarding the Child (e.g., "Have you been here already?," "What's your favorite toy?"), and asks about a female colleague (i.e., the Experimenter) that he is looking for. The episode ends with the Stranger saying "Hello!" to both the Child and Experimenter.	Baseline (segment 1: Child + Experimenter, briefing) Reactivity (segment 2: Child alone) (segment 3: Child + Stranger, approach) (segment 4: Child alone) Recovery (segment 5: Child + Experimenter, "hello!")
2	Perfect circle	The Experimenter asks the Child to draw a perfect circle, without giving guidelines about how a perfect circle should be. For about 3 min the Experimenter says "No, this is not perfect. Try again" to every attempt made by the Child. The episode ends with the Child making a happy face out of the last "perfect" circle.	Baseline (segment 1: Child + Experimenter, briefing) Reactivity (segment 2: Child + Experimenter, drawing) Recovery (segment 3: Child + Experimenter, happy face)
3	No sticker left	The Experimenter asks the Child to choose one among different cartoon-related stickers. Once the Child chooses a sticker, the Experimenter says: "Ok, now wait here, I need to give the other stickers to other children around here. I'll be back with yours in a while." When the Experimenter comes back, she has no stickers left for the Child. The episode ends with the Experimenter finding the Child's sticker and giving it to him.	Baseline (segment 1: Child + Experimenter, choosing) (segment 2: Child + Experimenter, describing) Reactivity (segment 3: Child alone) (segment 4: Child + Experimenter, no sticker) (segment 5: Child alone) Recovery (segment 6: Child + Experimenter, sticker back)
4	Transparent box	The Child is given a transparent box which contains an attractive toy. However, the box is closed with a lock and the Child is given the wrong key to open it. After 4 min of attempts alone, the Experimenter comes back into the room with the right key and the Child can play with the toy.	Baseline (segment 1: Child + Experimenter, briefing) Reactivity (segment 2: Child alone, attempts) Recovery (segment 3: Child + Experimenter, playing)

hold expertise in the coding of infants and children's behavior. Each coder independently watched 10 pilot applications of the PRES and provided a list of potential codes of the child's behavior every 10 s. Coders were asked to annotate the exact timing at which the selected behavioral codes occurred in order to facilitate the subsequent consensus discussion. In order to allow the coders to produce an adequate number of potential codes, no theoretical nor methodological limitations were imposed and descriptive rather than conceptual language was encouraged.

Consensus Procedure for Univocal Coding System

Subsequently, the first set of codes underwent a consensus discussion between the researchers. Specifically, overlapping codes identified by both LP and RMC survived the first screening, whereas codes present in the list of only one coder were discussed in *ad hoc* meetings. When a code was proposed only by one coder, different scenarios were possible. First, coders checked for potential overlapping of the code with previously identified and accepted codes. Second, if not overlapping, the time-frame in which the code occurred was reviewed by both coders together with a senior researcher (author RM). After this consensus process, the code was either suppressed or included. The final set of selected codes is reported in **Table 4**. They were separated in general codes (**Table 4A**) and specific codes (**Table 4B**), on the basis of their occurrence throughout the entire PRES procedure or limitedly to specific episodes, respectively.

Computing of the ESR General and Specific Indexes

The coding of the PRES procedure is micro-analytical (i.e., 10-s epochs). Every 10 s, the coders have to attribute a level of each general and specific code. A series of algorithms was developed in order to obtain general and specific indexes starting from general and specific codes, respectively. Prior to the computation of general and specific indexes of ESR, each code was weighted on the actual duration (i.e., number of epochs) of each episode's phase (i.e., baseline, reactivity, and recovery). As such, every code is expressed as a proportion ranging from 0 (never occurring) to 1 (always occurring).

General indexes. General codes are resumed into two general indexes: active engagement and stress level. General indexes are computed separately for the baseline, reactivity and recovery phases of each PRES episode.

Active engagement is calculated as the sum of the following general codes: sit (activity level), positive and neutral (emotional state), adult-directed and object-directed gaze (gaze direction). As such, active engagement ranges from 0 (all the included codes never occur throughout the specific phase of the episode) to 4 (all the included codes occur during every epoch of the specific phase of the episode).

Stress level is computed as the sum of the following general codes: stand (activity level), negative (emotional state), hands and mouth movements (peripheral movements), gaze aversion (gaze direction). As such, stress level ranges from 0 (all the included codes never occur throughout the specific phase of the episode)

TABLE 4 | List of general **(A)** and specific **(B)** codes.

(A)

Codes	Levels	Description
Activity level	Sit	The child sits on the chair
	Stand	The child stands up in front of the table
	Exploring	The child is exploring the environment, moving around the room
Emotional state	Positive	Facial, body or vocal expressions of positive affect
	Neutral	Absence of clear signs of positive or negative affect
	Negative	Facial, body or vocal expressions of negative affect
Gaze direction	Object-directed	The child maintains eye-contact with a task-related object
	Adult-directed	The child maintains eye-contact with a task-related adult
	Gaze aversion	The child averts eye-contact with task-related objects/adults
Peripheral movements	Hands	Subtle hands movements (e.g., grasping, rubbing, playing nervously) which are not associated with communications, playing, task-related activities or exploring
	Mouth	Subtle mouth movements (e.g., grimaces, biting the lips, tongue movements, stretching the lips) which are not associated with communications, playing, task-related activities or exploring
	Legs	Leg movements which are not embedded in general movements of the child (e.g., not related to postural changes)
	Arms	Arm movements which are not embedded in general movements of the child (e.g., not related to postural changes)

(B)

Episodes	Codes	Specific index
1	Child waits for the experimenter when she exits	Control
	Child moves toward the door when the experimenter is out of the room	Control
	Child asks about the experimenter where she is going when she exits	Control
	Child asks where the experimenter is to the stranger	Control
	Child checks for the experimenter out of the door	Control
	Child moves away from the stranger	Fear
	Child opens the door to exit	Fear
	Child does not respond to the stranger's questions	Inhibition
	Child does not describe the stranger to the experimenter	Inhibition
	Child needs help to describe the stranger to the experimenter	Inhibition
	Child has no memory of the stranger	Inhibition
2	Child expresses anger or protest (verbal and non-verbal)	Anger
	Child asks how the perfect circle should be drawn	Control
	Child looks at the experimenter instead of drawing	Control
	Child looks at the experimenter waiting to receive a feedback	Control
	Child asks to turn the sheet	Control
	Child does not draw the circle	Inhibition
	Child does not draw the smiling face	Inhibition
	Child expresses sadness or resignation	Sadness
3	Child expresses anger or protest (verbal and non-verbal)	Anger
	Child asks about the experimenter where she is going when she exits	Control
	Child waits for the experimenter when she exits	Control
	Child opens the door to exit	Fear
	Child asks about stickers' destiny	Control
	Child does not choose any sticker	Inhibition
	Child does not describe the chosen sticker	Inhibition
	Child does not accept back the sticker	Inhibition
	Child expresses sadness or resignation (verbal and non-verbal)	Sadness
4	Child expresses anger or protest (verbal and non-verbal)	Anger

(Continued)

TABLE 4 | Continued

(B)

Episodes	Codes	Specific index
	Child asks about the experimenter where she is going when she exits	Control
	Child opens the door to exit	Fear
	Child gives up at attempting to open the box	Inhibition
	Child expresses sadness or resignation (verbal and non-verbal)	Sadness
	Child does not play with the toy	Inhibition
	Child plays with the toy in aggressive ways	Anger

to 4 (all the include codes occur during every epoch of the specific phase of the episode).

Specific indexes. The development of specific indexes of ESR followed a qualitative labeling procedure. The two independent coders attached a one-word label to each specific code. For instance, the specific code "The child moves away from the stranger" was labeled *Fear*; the specific code "No response to stranger" was labeled *Inhibition*. Furthermore, disagreement between the two coders' labeling was discussed and resolved. Five labels were produced, corresponding to the final set of five specific indexes: *anger*, *control*, *fear*, *inhibition*, and *sadness*.

Each specific index is computed as the sum of the related specific codes (weighted on the actual duration of each episode's phase, i.e., number of epochs). Due to the different types of emotional stress elicited by the PRES episodes, specific indexes cannot be scored for all episodes. Consistently, as varying theoretical ranges apply to different episodes, they are meant to be standardized, with mean = 0 and standard deviation = 1.

ANTICIPATED RESULTS

Proofs of Reliability

Reliability of the coding system will be assessed using multiple methods. First, inter-coder reliability will be measured according to Cohen's k coefficient and percentage agreement. Second, test–retest reliability will be assessed according to Cronbach's alpha. Finally, a confirmatory factorial analysis will be used to verify if the theoretically aggregated general and specific indexes are supported by statistic clustering.

Strengths and Limitations

The PRES presents specific advantages and potential strengths when compared to available procedures in literature. It has been developed within a well-defined theoretical framework (i.e., infant research tradition) in which both reactivity and recovery are included as two adaptive steps of ESR. Second, the coding system is micro-analytical, which allows researchers to obtain fine-grained information on children's behavioral responses to emotionally challenging conditions with no need of abstraction or global ratings. Moreover, the procedural and descriptive definition of each code is meant to facilitate the agreement among independent coders and to limit the risk of subjective interpretations. Third, it provides two

levels of information on children ESR, including general indexes of negative emotionality and interactive engagement as well as specific indexes of different emotional responses to a stressful condition. Fourth, the procedure includes different challenging conditions which represent the main sources of emotional stress in preschool aged children. As such, the PRES is meant to be a multi-faceted and comprehensive assessment of ESR in preschoolers with a unified theoretical background, which limits the need of integrating different procedures or protocols in future research in the field.

Nonetheless, potential limitations exist. First, the PRES has been developed as a laboratory procedure. As such, its application to naturalistic settings (e.g., primary care, home environment) may require adaptations. Second, the micro-analytical nature of the PRES coding system is time- and resource-consuming. As such, despite the PRES is well-suited for research purposes, its application to clinical settings in which more rapid assessments are needed to sustain health-related decisions is limited. Third, the PRES does not include a direct evaluation of ESR in the context of peer-related stress. Despite the Stickers episode elicits stress that is related to the unequal treatment of the subject compared to non-present peers, there are no PRES episodes during which the child is required to interact with other children of the same age. Peer relationships involve many different dimensions (e.g., competition and cooperation) which are only limitedly linked with ESR and which require specific observational methods.

Ethics and Dissemination

The study has been approved by the Ethical Committee of the Scientific Institute IRCCS Eugenio Medea, Bosisio Parini (Italy). The PRES is intended to provide micro-analytical, intensive and rich information on the socio-emotional stress regulation of preschool children and the quantitative nature of this observational procedure is suitable within cross-sectional studies comparing low- and high-risk children as well as within longitudinal studies assessing the long-term effects of early adversities on emotional development. For instance, the PRES procedure and the related coding system is currently being used in a prospective longitudinal research project on the epigenetic correlates of early adversity exposure in prematurity (the preschool-age phase is currently ongoing and data on the infancy phase are published: Provenzi et al., 2015b;

Montirosso et al., 2016a,b). In this study, a clinical group of preschool aged children born preterm, which are known to be at risk of altered ESR (Montagna and Nosarti, 2016), will be compared to a control group of full-term preschool peers at 4.5 years. The comparison of the response to the PRES procedure between the two groups will also serve as a preliminary validation of the capacity of this laboratory assessment paradigm to depict difficulties in ESR in at-risk preschool aged children. Further methodological steps of the PRES validation and its application to larger samples of low- and high-risk preschool children will be reported in future conference presentations and peer-reviewed journals.

AUTHOR CONTRIBUTIONS

LP and RC developed the first version of the manual and the coding system. LP wrote the first draft of the present manuscript. GSdM contributed to final manuscript editing and English editing. ML and RM approved the final version of the manuscript. RM contributed to the refining of the manual, the coding system and the final paper.

REFERENCES

Adamson, L. B., and Frick, J. E. (2003). The still face: a history of a shared experimental paradigm. *Infancy* 4, 451–473. doi: 10.1207/S15327078IN 0404_01

Adrian, M., Zeman, J., and Veits, G. (2011). Methodological implications of the affect revolution: a 35-year review of emotion regulation assessment in children. *J. Exp. Child Psychol.* 110, 171–197. doi: 10.1016/j.jecp.2011.03.009

Ahnert, L., Gunnar, M. R., Lamb, M. E., and Barthel, M. (2004). Transition to child care: associations with infant-mother attachment, infant negative emotion, and cortisol elevations. *Child Dev.* 75, 639–650. doi: 10.1111/j.1467-8624.2004. 00698.x

Aldao, A., Gee, D. G., De Los Reyes, A., and Seager, I. (2016). Emotion regulation as a transdiagnostic factor in the development of internalizing and externalizing psychopathology: current and future directions. *Dev. Psychopathol.* 28(4 Pt 1), 927–946. doi: 10.1017/S0954579416000638

Bariola, E., Gullone, E., and Hughes, E. K. (2011). Child and adolescent emotion regulation: the role of parental emotion regulation and expression. *Clin. Child Fam. Psychol. Rev.* 14, 198–212. doi: 10.1007/s10567-011-0092-5

Beebe, B., and Steele, M. (2013). How does microanalysis of mother-infant communication inform maternal sensitivity and infant attachment? *Attach. Hum. Dev.* 15, 583–602. doi: 10.1080/14616734.2013.841050

Beeghly, M., Fuertes, M., Liu, C. H., Delonis, M. S., and Tronick, E. (2011). "Maternal sensitivity in dyadic context: mutual regulation, meaning-making, and reparation," in *Maternal Sensitivity: A Scientific Foundation for Practice*, eds D. W. Davis and M. C. Logsdon (Hauppauge, NY: Nova Science Publishers, Inc.).

Beeghly, M., and Tronick, E. (2011). Early resilience in the context of parentinfant relationships: a social developmental perspective. *Curr. Probl. Pediatr. Adolesc. Health Care* 41, 197–201. doi: 10.1016/j.cppeds.2011.02.005

Berking, M., and Wupperman, P. (2012). Emotion regulation and mental health. *Curr. Opin. Psychiatry* 25, 128–134. doi: 10.1097/YCO.0b013e3283503669

Blankson, A. N., Weaver, J. M., Leerkes, E. M., O'Brien, M., Calkins, S. D., and Marcovitch, S. (2017). Cognitive and emotional processes as predictors of a successful transition into school. *Early Educ. Dev.* 28, 1–20. doi: 10.1080/ 10409289.2016.1183434

Bornstein, M. H., and Manian, N. (2013). Maternal responsiveness and sensitivity reconsidered: some is more. *Dev. Psychopathol.* 25, 957–971. doi: 10.1017/ S0954579413000308

Braungart-Rieker, J., Garwood, M. M., Powers, B. P., and Notaro, P. C. (1998). Infant affect and affect regulation during the still-face paradigm with mothers and fathers: the role of infant characteristics and parental sensitivity. *Dev. Psychol.* 34, 1428–1437. doi: 10.1037/0012-1649.34.6.1428

Buss, K. A., and Goldsmith, H. H. (1998). Fear and anger regulation in infancy: effects on the temporal dynamics of affective expression. *Child Dev.* 69, 359–374. doi: 10.1111/j.1467-8624.1998.tb06195.x

Buss, K. A., and Kiel, E. J. (2011). Do maternal protective behaviors alleviate toddlers' fearful distress? *Int. J. Behav. Dev.* 35, 136–143. doi: 10.1177/ 0165025410375922

Cisler, J. M., and Olatunji, B. O. (2012). Emotion regulation and anxiety disorders. *Curr. Psychiatry Rep.* 14, 182–187. doi: 10.1007/s11920-012-0262-2

Cole, P. M., Teti, L. O., and Zahn-Waxler, C. (2003). Mutual emotion regulation and the stability of conduct problems between preschool and early school age. *Dev. Psychopathol.* 15, 1–18. doi: 10.1017/S0954579403000014

Compas, B. E., Jaser, S. S., Dunbar, J. P., Watson, K. H., Bettis, A. H., Gruhn, M. A., et al. (2014). Coping and emotion regulation from childhood to early adulthood: points of convergence and divergence. *Aust. J. Psychol.* 66, 71–81. doi: 10.1111/ajpy.12043

DiCorcia, J. A., and Tronick, E. D. (2011). Quotidian resilience: exploring mechanisms that drive resilience from a perspective of everyday stress and coping. *Neurosci. Biobehav. Rev.* 35, 1593–1602. doi: 10.1016/j.neubiorev.2011. 04.008

Dvir, Y., Ford, J. D., Hill, M., and Frazier, J. A. (2014). Childhood maltreatment, emotional dysregulation, and psychiatric comorbidities. *Harv. Rev. Psychiatry* 22, 149–161. doi: 10.1097/HRP.0000000000000014

Dyson, M. W., Olino, T. M., Durbin, C. E., Goldsmith, H. H., Bufferd, S. J., Miller, A. R., et al. (2015). The structural and rank-order stability of temperament in young children based on a laboratory-observational measure. *Psychol. Assess.* 27, 1388–1401. doi: 10.1037/pas0000104

Eisenberg, N., Spinrad, T. L., and Eggum, N. D. (2010). Emotion-related self-regulation and its relation to children's maladjustment. *Annu. Rev. Clin. Psychol.* 6, 495–525. doi: 10.1146/annurev.clinpsy.121208.131208

Ekas, N. V., Haltigan, J. D., and Messinger, D. S. (2013). The dynamic still-face effect: Do infants decrease bidding over time when parents are not responsive? *Dev. Psychol.* 49, 1027–1035. doi: 10.1037/a0029330

Feldman, R. (2007). Mother-infant synchrony and the development of moral orientation in childhood and adolescence: direct and indirect mechanisms of developmental continuity. *Am. J. Orthopsychiatry* 77, 582–597. doi: 10.1037/ 0002-9432.77.4.582

Ford, B. Q., Mauss, I. B., Troy, A. S., Smolen, A., and Hankin, B. (2014). Emotion regulation moderates the risk associated with the 5-HTT gene and stress in children. *Emotion* 14, 930–939. doi: 10.1037/a0036835

Fox, N. A., and Calkins, S. D. (2003). The development of self-control of emotion: intrinsic and extrinsic influences. *Motiv. Emot.* 27, 7–26. doi: 10.1023/A: 1023622324898

Gagne, J. R., Van Hulle, C. A., Aksan, N., Essex, M. J., and Goldsmith, H. H. (2011). Deriving childhood temperament measures from emotion-eliciting behavioral episodes: scale construction and initial validation. *Psychol. Assess.* 23, 337–353. doi: 10.1037/a0021746

Gartstein, M. A., and Marmion, J. (2008). Fear and positive affectivity in infancy: convergence/discrepancy between parent-report and laboratory-based indicators. *Infant Behav. Dev.* 31, 227–238. doi: 10.1016/j.infbeh.2007.10.012

Gilliom, M., Shaw, D. S., Beck, J. E., Schonberg, M. A., and Lukon, J. L. (2002). Anger regulation in disadvantaged preschool boys: strategies, antecedents, and the development of self-control. *Dev. Psychol.* 38, 222–235. doi: 10.1037/0012-1649.38.2.222

Goldsmith, H. H., Buss, A. H., Plomin, R., Rothbart, M. K., Thomas, A., Chess, S., et al. (1987). Roundtable: What is temperament? Four approaches. *Child Dev.* 58, 505–529. doi: 10.2307/1130527

Goldsmith, H. H., Reilly, J., Lemery, K. S., Longley, S., and Prescott, A. (1993). *Preliminary Manual for the Preschool Laboratory Temperament Assessment Battery*. Technical Report. Madison, WI: University of Wisconsin.

Gunnar, M. R., Talge, N. M., and Herrera, A. (2009). Stressor paradigms in developmental studies: what does and does not work to produce mean increases in salivary cortisol. *Psychoneuroendocrinology* 34, 953–967. doi: 10.1016/j.psyneuen.2009.02.010

Hsu, H.-C., and Jeng, S.-F. (2008). Two-month-olds' attention and affective response to maternal still face: a comparison between term and preterm infants in Taiwan. *Infant Behav. Dev.* 31, 194–206. doi: 10.1016/j.infbeh.2007.10.008

Hughes, C., Oksanen, H., Taylor, A., Jackson, J., Murray, L., Caspi, A., et al. (2002). "I'm gonna beat you!' SNap!: an observational paradigm for assessing young children's disruptive behaviour in competitive play. *J. Child Psychol. Psychiatry* 43, 507–516. doi: 10.1016/j.hrthm.2016.03.039

Huh, H. J., Kim, K. H., Lee, H.-K., and Chae, J.-H. (2017). The relationship between childhood trauma and the severity of adulthood depression and anxiety symptoms in a clinical sample: the mediating role of cognitive emotion regulation strategies. *J. Affect. Disord.* 213, 44–50. doi: 10.1016/j.jad.2017.02.009

Kaitz, M., Maytal, H. R., Devor, N., Bergman, L., and Mankuta, D. (2010). Maternal anxiety, mother-infant interactions, and infants' response to challenge. *Infant Behav. Dev.* 33, 136–148. doi: 10.1016/j.infbeh.2009.12.003

Kim, S., and Kochanska, G. (2017). Relational antecedents and social implications of the emotion of empathy: evidence from three studies. *Emotion* 17, 981–992. doi: 10.1037/emo0000297

Kochanska, G., Aksan, N., Penney, S. J., and Doobay, A. F. (2007). Early positive emotionality as a heterogeneous trait: implications for children's self-regulation. *J. Pers. Soc. Psychol.* 93, 1054–1066. doi: 10.1037/0022-3514.93.6.1054

Kochanska, G., and Knaack, A. (2003). Effortful control as a personality characteristic of young children: antecedents, correlates, and consequences. *J. Pers.* 71, 1087–1112. doi: 10.1111/1467-6494.7106008

Kopala-Sibley, D. C., Dougherty, L. R., Dyson, M. W., Laptook, R. S., Olino, T. M., Bufferd, S. J., et al. (2017). Early childhood cortisol reactivity moderates the effects of parent-child relationship quality on the development of children's temperament in early childhood. *Dev. Sci.* 20:e12378. doi: 10.1111/desc.12378

Langerock, N., van Hanswijck de Jonge, L., Bickle Graz, M., Hüppi, P. S., Borradori Tolsa, C., and Barisnikov, K. (2013). Emotional reactivity at 12 months in very preterm infants born at. *Infant Behav. Dev.* 36, 289–297. doi: 10.1016/j.infbeh.2013.02.006

Lesch, K. P. (2011). When the serotonin transporter gene meets adversity: the contribution of animal models to understanding epigenetic mechanisms in affective disorders and resilience. *Curr. Top. Behav. Neurosci.* 7, 251–280. doi: 10.1007/7854_2010_109

Linden, W., Earle, T. L., Gerin, W., and Christenfeld, N. (1997). Physiological stress reactivity and recovery: conceptual siblings separated at birth? *J. Psychosom. Res.* 42, 117–135. doi: 10.1016/S0022-3999(96)00240-1

Malone, S. M., Gunnar, M. R., and Risch, R. O. (1985). Adrenocortical and behavioral responses to limb restraint in human neonates. *Dev. Psychobiol.* 18, 435–446. doi: 10.1002/dev.420180508

McLaughlin, K. A., Hatzenbuehler, M. L., Mennin, D. S., and Nolen-Hoeksema, S. (2011). Emotion dysregulation and adolescent psychopathology: a prospective study. *Behav. Res. Ther.* 49, 544–554. doi: 10.1016/j.brat.2011.06.003

Melnick, S. M., and Hinshaw, S. P. (2000). Emotion regulation and parenting in AD/HD and comparison boys: linkages with social behaviors and peer preference. *J. Abnorm. Child Psychol.* 28, 73–86. doi: 10.1023/A:1005174102794

Mesman, J., van IJzendoorn, M. H., and Bakermans-Kranenburg, M. J. (2009). The many faces of the Still-Face Paradigm: a review and meta-analysis. *Dev. Rev.* 29, 120–162. doi: 10.1016/j.dr.2009.02.001

Montagna, A., and Nosarti, C. (2016). Socio-emotional development following very preterm birth: pathways to psychopathology. *Front. Psychol.* 7:80. doi: 10.3389/fpsyg.2016.00080

Montirosso, R., Borgatti, R., Trojan, S., Zanini, R., and Tronick, E. (2010). A comparison of dyadic interactions and coping with still-face in healthy pre-term and full-term infants. *Br. J. Dev. Psychol.* 28, 347–368. doi: 10.1348/026151009X416429

Montirosso, R., Casini, E., Provenzi, L., Putnam, S. P., Morandi, F., Fedeli, C., et al. (2015a). A categorical approach to infants' individual differences during the Still-Face paradigm. *Infant Behav. Dev.* 38, 67–76. doi: 10.1016/j.infbeh.2014.12.015

Montirosso, R., Provenzi, L., Fumagalli, M., Sirgiovanni, I., Giorda, R., Pozzoli, U., et al. (2016a). Serotonin transporter gene (*SLC6A4*) methylation associates with neonatal intensive care unit stay and 3-month-old temperament in preterm infants. *Child Dev.* 87, 38–48. doi: 10.1111/cdev.12492

Montirosso, R., Provenzi, L., Giorda, R., Fumagalli, M., Morandi, F., Sirgiovanni, I., et al. (2016b). SLC6A4 promoter region methylation and socio-emotional stress response in very preterm and full-term infants. *Epigenomics* 8, 895–907. doi: 10.2217/epi-2016-0010

Montirosso, R., Provenzi, L., Tavian, D., Morandi, F., Bonanomi, A., Missaglia, S., et al. (2015b). Social stress regulation in 4-month-old infants: contribution of maternal social engagement and infants' 5-HTTLPR genotype. *Early Hum. Dev.* 91, 173–179. doi: 10.1016/j.earlhumdev.2015.01.010

Morris, A. S., Silk, J. S., Morris, M. D. S., Steinberg, L., Aucoin, K. J., and Keyes, A. W. (2011). The influence of mother-child emotion regulation strategies on children's expression of anger and sadness. *Dev. Psychol.* 47, 213–225. doi: 10.1037/a0021021

Pauli-Pott, U., Friedl, S., Hinney, A., and Hebebrand, J. (2009). Serotonin transporter gene polymorphism (5-HTTLPR), environmental conditions, and developing negative emotionality and fear in early childhood. *J. Neural Transm.* 116, 503–512. doi: 10.1007/s00702-008-0171-z

Pisani, A. R., Wyman, P. A., Petrova, M., Schmeelk-Cone, K., Goldston, D. B., Xia, Y., et al. (2013). Emotion regulation difficulties, youth–adult relationships, and suicide attempts among high school students in underserved communities. *J. Youth Adolesc.* 42, 807–820. doi: 10.1007/s10964-012-9884-2

Planalp, E. M., and Braungart-Rieker, J. M. (2015). Trajectories of regulatory behaviors in early infancy: determinants of infant self-distraction and self-comforting. *Infancy* 20, 129–159. doi: 10.1111/infa.12068

Provenzi, L., Borgatti, R., Menozzi, G., and Montirosso, R. (2015a). A dynamic system analysis of dyadic flexibility and stability across the Face-to-Face Still-Face procedure: application of the state space grid. *Infant Behav. Dev.* 38, 1–10. doi: 10.1016/j.infbeh.2014.10.001

Provenzi, L., Fumagalli, M., Sirgiovanni, I., Giorda, R., Pozzoli, U., Morandi, F., et al. (2015b). Pain-related stress during the Neonatal Intensive Care Unit stay and SLC6A4 methylation in very preterm infants. *Front. Behav. Neurosci.* 9:99. doi: 10.3389/fnbeh.2015.00099

Provenzi, L., Giusti, L., and Montirosso, R. (2016). Do infants exhibit significant cortisol reactivity to the Face-to-Face Still-Face paradigm? A narrative review and meta-analysis. *Dev. Rev.* 42, 34–55. doi: 10.1016/j.dr.2016.07.001

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., and Fisher, P. (2001). Investigations of temperament at three to seven years: the Children's Behavior Questionnaire. *Child Dev.* 72, 1394–1408. doi: 10.1111/1467-8624.00355

Rothbart, M. K., Posner, M. I., and Kieras, J. (2006). "Temperament, attention, and the development of self-regulation," in *Blackwell Handbook of Early Childhood Development*, eds K. McCartney and D. Phillips (Oxford: Blackwell Publishing Ltd.), 338–357. doi: 10.1002/9780470757703.ch17

Rothbart, M. K., Sheese, B. E., Rueda, M. R., and Posner, M. I. (2011). Developing mechanisms of self-regulation in early life. *Emot. Rev.* 3, 207–213. doi: 10.1177/1754073910387943

Saudino, K. J., and Micalizzi, L. (2015). Emerging trends in behavioral genetic studies of child temperament. *Child Dev. Perspect.* 9, 144–148. doi: 10.1111/cdep.12123

Shanahan, M., Roberts, J., Hatton, D., Reznick, J., and Goldsmith, H. (2008). Early temperament and negative reactivity in boys with fragile X syndrome. *J. Intellect. Disabil. Res.* 52, 842–854. doi: 10.1111/j.1365-2788.2008.01074.x

Stifter, C. A., and Braungart, J. M. (1995). The regulation of negative reactivity in infancy: function and development. *Dev. Psychol.* 31, 448–455. doi: 10.1037/0012-1649.31.3.448

Szentágotai-Tãtar, A., and Miu, A. C. (2016). Individual differences in emotion regulation, childhood trauma and proneness to shame and guilt in adolescence. *PLOS ONE* 11:e0167299. doi: 10.1371/journal.pone.0167299

Thompson, R. A. (1994). Emotion regulation: a theme in search of definition. *Monogr. Soc. Res. Child Dev.* 59, 25–52. doi: 10.2307/1166137

Tronick, E., Als, H., Adamson, L., Wise, S., and Brazelton, T. B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *J. Am. Acad. Child Psychiatry* 17, 1–13. doi: 10.1016/S0002-7138(09)62273-1

Tsigos, C., Kyrou, I., Kassi, E., Chrousos, G. P., and De, Groot LJ (2000). *Stress, Endocrine Physiology and Pathophysiology*, eds G. Chrousos, K. Dungan, K. R. Feingold, A. Grossman, J. M. Hershman, C. Koch, et al. (South Dartmouth, MA: MDText.com, Inc.).

Waider, J., Araragi, N., Gutknecht, L., and Lesch, K.-P. (2011). Tryptophan hydroxylase-2 (TPH2) in disorders of cognitive control and emotion regulation: a perspective. *Psychoneuroendocrinology* 36, 393–405. doi: 10.1016/j.psyneuen.2010.12.012

Weinberg, M. K., and Tronick, E. Z. (1998). Emotional characteristics of infants associated with maternal depression and anxiety. *Pediatrics* 102(5 Suppl. E), 1298–1304.

Yato, Y., Kawai, M., Negayama, K., Sogon, S., Tomiwa, K., and Yamamoto, H. (2008). Infant responses to maternal still-face at 4 and 9 months. *Infant Behav. Dev.* 31, 570–577. doi: 10.1016/j.infbeh.2008.07.008

Zimmer-Gembeck, M. J., Webb, H. J., Pepping, C. A., Swan, K., Merlo, O., Skinner, E. A., et al. (2015). Review: Is parent–child attachment a correlate of children's emotion regulation and coping? *Int. J. Behav. Dev.* doi: 10.1177/0165025415618276 [Epub ahead of print].

Zimmermann, L. K., and Stansbury, K. (2004). The influence of emotion regulation, level of shyness, and habituation on the neuroendocrine response of three-year-old children. *Psychoneuroendocrinology* 29, 973–982. doi: 10.1016/j.psyneuen.2003.09.003

Zmyj, N., Schölmerich, A., and Daum, M. M. (2017). The relationship between attention and deferred imitation in 12-month-old infants. *Infant Behav. Dev.* 48(Pt B), 175–183. doi: 10.1016/j.infbeh.2017.02.001

Consumer Neuroscience-Based Metrics Predict Recall, Liking and Viewing Rates in Online Advertising

Jaime Guixeres[1]*, Enrique Bigné[2], Jose M. Ausín Azofra[1], Mariano Alcañiz Raya[1], Adrián Colomer Granero[1], Félix Fuentes Hurtado[1] and Valery Naranjo Ornedo[1]

[1] *Instituto de Investigación e Innovación en Bioingeniería, Universidad Politécnica de València, València, Spain,*
[2] *Departamento de Comercialización e Investigación de Mercados, Facultad de Economía, Universitat de València, València, Spain*

The purpose of the present study is to investigate whether the effectiveness of a new ad on digital channels (YouTube) can be predicted by using neural networks and neuroscience-based metrics (brain response, heart rate variability and eye tracking). Neurophysiological records from 35 participants were exposed to 8 relevant TV Super Bowl commercials. Correlations between neurophysiological-based metrics, ad recall, ad liking, the ACE metrix score and the number of views on YouTube during a year were investigated. Our findings suggest a significant correlation between neuroscience metrics and self-reported of ad effectiveness and the direct number of views on the YouTube channel. In addition, and using an artificial neural network based on neuroscience metrics, the model classifies (82.9% of average accuracy) and estimate the number of online views (mean error of 0.199). The results highlight the validity of neuromarketing-based techniques for predicting the success of advertising responses. Practitioners can consider the proposed methodology at the design stages of advertising content, thus enhancing advertising effectiveness. The study pioneers the use of neurophysiological methods in predicting advertising success in a digital context. This is the first article that has examined whether these measures could actually be used for predicting views for advertising on YouTube.

Keywords: neuromarketing, YouTube, artificial neural networks, eye tracking, heart rate variability, brain response

INTRODUCTION

Advertising effectiveness is still challenging academics and practitioners. Neuroimaging and physiological measurement tools are becoming popular within marketing (Daugherty et al., 2016). Their primary uses are related to unconscious measures based on eye movement, heart rate and brain activity, among others (see Venkatraman et al., 2015 for more details). Such tools aim to provide better understanding of the impact of affect and cognition on memory (Vecchiato et al., 2013). Furthermore, neurophysiological methods can capture the dynamics of television commercials content because they provide continuous data unlike traditional measures, such as interviews and surveys that only reflect a global indicator for every commercial.

Global expenditure on media has been rising over the years and digital advertising is the fastest-growing category (McKinsey, 2015). Marketers are still calling for accurate assessments about advertisement effectiveness and about the return on advertising expenditure (McAlister et al., 2016). Nowadays, Internet advertising has evolved dramatically and a platform like YouTube is a good example on how to reach viewers at a global scale.

Advertisers pursue the attention of viewers and seek ad recall, brand recall and positive emotions. If this occurs, ads will be stored in the viewers' long-term memory. Humans may not remember each advertisement they have been exposed to, but neuroscience techniques can detect conditions that lead to the memorization of advertising.

The usage of neuro metrics in measuring advertising effectiveness overcome some of the weaknesses associated with traditional measures (Varan et al., 2015). Despite the benefits of using such measurements, the key question is the choosing of the variables of advertising effectiveness. Among the different types of effects pursue by advertising (see Moriarty et al., 2012), three types of effects have been considered. First, perception, and particularly exposure to the ad is recognized as the first step in any evaluation process. In this study, we adopt the online views as the measure of advertising effectiveness of online ads; secondly, the emotional dimension is typically used in evaluating the effects of advertising, thus we adopt the liking as an emotional metric; lastly, the cognition effect of advertising is measured through ad recall.

Neurophysiological methods offer richer data than self-report measurements of particular interest in advertising research. Firstly, physiological measurements of emotion allow researchers to analyze emotional activity without cognitive bias. Secondly, neurophysiological methods provide instant and continuous data that allow researchers to decompose the data analysis into small pieces of study. Lastly, physiological measurements typically offer a myriad of metrics. In the present study several metrics are compared and two new metrics derived from eye tracking (ET) are also proposed: "number of quadrants per second" (Quad_sec) and "gaze brand effectiveness ratio" (Brand_ratio). However, physiological measures have their own limitations: a strong reliance on physiological data to measure emotions leaves room for misinterpretation of physiological noise (e.g., natural changes in body status) and burdens researchers with the difficult task of attributing specific physiological changes (e.g., increase in heart rate) to complex and subjectively experienced emotions (e.g., hate, love, or fear).

Scholar research has recently adopted neurophysiological measures to better understand consumer responses to advertising (Astolfi et al., 2009). To the best of our knowledge, no articles in marketing have previously examined whether these measures could actually be transferred into real life views on advertising on YouTube. The work by Venkatraman et al. (2015) was one of the first pieces of research that tried to establish correlations between brand performance and physiological responses whilst viewers watched ads. So far it has been hard to gauge the number of viewers who will watch an ad, which in turn is one of the main objectives for marketers. However, digital platforms overcome this situation enabling researchers to measure consumer unconscious reactions to ads and the number of views.

Neurophysiological methods to measure advertising effectiveness are becoming popular including consolidated tools such as ET and facial reader (Wedel and Pieters, 2014), EGG (Ohme et al., 2010) and more recently sophisticated tools such as fMRI (Venkatraman et al., 2015; Couwenberg et al., 2017).

New extensions to sales are emerging, Thus, biomarkers can, to some extent, predict sales figures (see Kühn et al., 2016) and even in virtual reality experiments have been developed (see Bigné et al., 2016).

This paper aims to answer whether neurophysiological methods contribute anything beyond traditional methods in predicting ad success in a digital context. Specific research goals are listed below. Firstly, to analyze whether three of the most cited neurophysiological and behavioral techniques: electroencephalogram (EEG), heart rate variability (HRV) and ET correlate with common cognitive states typically used in advertising research (e.g., liking and recall measures) (Morin, 2011; Ruanguttamanun, 2014; dos Santos et al., 2015). Secondly, we aim to explain whether the variance in the number of views on a brand's official YouTube channel is related to any of the neurophysiological measures and their metrics.

A study to assess subjects' responses to nine 30-s online television ads was conducted. Data gathered from 47 subjects were split into six datasets based on three conditions: recall (RMB) vs. no-recall (FRG), liking (LIKE vs. DISLIKE) and Internet views (>5M vs. <5M). All the metrics extracted from physiological and behavioral responses were compared and correlated between these groups.

The contributions of this paper are listed hereafter. First, we show how different metrics from three neurophysiological devices correlate in an attempt to select the most accurate ones for digital commercials. Second, Artificial Neural Networks (ANN), using biometric data can predict digital views of ads, hence common physiological patterns related to unconscious responses can predict when an ad is going to be remembered or liked. Finally, two new metrics are proposed to measure advertising effectiveness in digital advertising, which show a high level of accuracy in predicting digital views. In this paper, we build on the later works that reviewed Super Bowl ads with ET and heart rate (Christoforou et al., 2015), and brain response (Deitz et al., 2016), but adding for the first time joined in neural networks models, three of the most employed signals in advertisement research (EEG, HRV and ET) and proposing two novel metrics based only on eye-tracking data that predicts viewer's preferences in video advertisements.

The rest of the paper is organized as follows. Firstly, we provide a brief literature review of advertising research effectiveness and we introduce neurophysiological methods related to ad recall and ad likeability. Secondly, we describe the experimental design and signal recording and processing techniques used to extract biometric data. Then, we describe the study results in three parts. In the first part, a comparison of the metrics from neurophysiological signals related to likeability of the ad and ad recall. In the second one, the correlations between these biometrics, the score given by participants in a poll (e.g., ACE_score) and the number of views on YouTube are exposed; and thirdly, by applying ANN to these biometric datasets, we predict the number of views on YouTube for each ad tested during the study. Finally, we discuss the contributions and implications for researchers and practitioners.

Established Methods in Advertising Research

Despite the diverse approaches used in advertising research (Vakratsas and Ambler, 1999), advertising success on ad execution has focused on traditional measures such as liking, excitability, and recall (Venkatraman et al., 2015). Acknowledging the established literature (Astolfi et al., 2009; Kim et al., 2014), this paper focuses on liking and recall as traditional measures. Online polls have been adopted in academic research as a valuable data source (see Strach et al., 2015). Based on a US national-representative Internet sample of 500 respondents, Ace Metrix has been providing advertising effectiveness scores since January 2009 and it is used in this study.

Advertising Research in a Digital Setting

Digital channels have changed advertising research dramatically and, as a result, a new paradigm is emerging (Ha, 2008; Bigné, 2016). One of the major gains is that analytics are available at ad level, including exposure measured through number of views and likeability through "likes."

Neurophysiological Tools in Advertising Research

This study focuses on three neurophysiological methods, aiming to collect data from different angles: eye movements, heart variability and brain responses.

Eye tracking is a well-established measure of visual attention (Wedel and Pieters, 2014; Venkatraman et al., 2015) to different stimuli, such as product choice (Guerreiro et al., 2015), static images (Mould et al., 2012), printed ads (Elsen et al., 2016) banner ads (Lee and Ahn, 2012) and videos of the Super Bowl (Christoforou et al., 2015).

Heart rate variability is the physiological phenomenon of variation in the time interval between heartbeats. It is measured by the variation in the beat-to-beat interval (Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996). This variability of the heart is related to activations of the sympathetic and parasympathetic systems of the autonomic nervous system. HRV provides an independent measure of attention (Lang et al., 1999) and it has been applied to television commercials (Acharya et al., 2006; Grandjean et al., 2008; Geisler et al., 2010; Bellman et al., 2013; Valenza et al., 2014; Venkatraman et al., 2015).

Electroencephalogram is an electrophysiological monitoring method to record the electrical activity of the brain. The relationship between affection, engagement and brain activation in frontal brain activity has been well documented in psychology and neuroscience research (Harmon-Jones et al., 2010; Khushaba et al., 2013). Emotional frontal asymmetry as hypothesized by Davidson (2004) has been applied to analyze commercials (Ohme et al., 2010; Vecchiato et al., 2011), including Super Bowl ads (Deitz et al., 2016) and advertising success (Venkatraman et al., 2015). Vecchiato et al. (2011, p. 582) showed that "activity" in the left-frontal cortex related to "pleasant" commercials and activity in the right-frontal cortex associated with "unpleasant" commercials.

Hypothesis Development

As briefly discussed earlier, ET, HRV and EEG provide measures of responses to advertising stimuli and might be related to ad performance. This study attempts to examine the relationship between three types of data from neurophysiological tools, ET, HRV and EEG, and three advertising variables typically used in advertising research, such as ad recall, ad likeability and ad views. Hypotheses will be anchored in three streams of research aiming to integrate them into a single approach: (i) theoretical advertising literature; (ii) online advertising; (iii) neurophysiological research related to advertising. An integrative approach is useful because neurophysiological primary data *per se* are non-meaningful for advertising research. Therefore, this type of data must be interpreted in relation to classic advertising assumptions in order to prove their validity. Most of the data gathered in this type of studies are based on a different methodological paradigm that derives from psychophysiology (Bolls et al., 2012).

1. Advertising research shows a positive relationship between recognition and attention. Donthu et al. (1993) found a positive relationship between attention and recall for outdoor advertising. In an ET study, Pieters et al. (2002) found that attention led to ad recognition. Baack et al. (2008) posit that recognition is considered an immediate measure of attention.
2. As discussed earlier, ET and some metrics from EEG can capture attention, hence if attention leads to recall, it could be argued that each neurophysiological tool can capture attention.

The Advertising Research Foundation's Copy Research Validity Project (CRVP) showed in the early nineties that advertising likability is the single best measure of effectiveness (Rossiter and Eagleson, 1994). Furthermore, likeability has been considered relevant and important in measuring commercial effectiveness in the ads aired in the Super Bowl, showing stable scores between 1990 and 1999 (Tomkovick et al., 2001). The positive influence of likeability has also been highlighted recently in online settings. Thus, likeability of online video ads has successfully linked to intention to share them (Shehu et al., 2016), which can be interpreted as a successful performance. Recent literature in neuromarketing also highlights a relationship between liking and HRV, ET measurements and fMRI signals (Venkatraman et al., 2015).

Online video platforms, such as YouTube, have been largely approached from the user-generated content perspective (see Smith et al., 2012). However, its dimension as a digital channel for watching commercials has almost been neglected, with some related exceptions (Verhellen et al., 2013). The number of views of each online video, including commercials, is available on YouTube, and it is commonly seen as a valid measure of its popularity. Given its social media nature, recent research is addressing two main fields of interest: the sources that drive views to a video and the preferred type of content. A recent study by Zhou et al. (2016) identified YouTube search and related video recommendation as the major view sources. In

adopting YouTube views, the age of the video and the potential replays must be considered. Research shows that user's preference seems relatively insensitive to the video's age (Cha et al., 2007). More recently, Chen et al. (2014) analyzed a lifetime model of online video popularity that features the following three main characteristics of interest here, adopting views as a potential variable in explaining our intended relationships: (i) views follow a Zipf distribution; (ii) replay percentage is very low; (iii) and only video content on news and sports are strongly dependent on age, with popularity being much less sensitive to age in music videos, which can be considered closer to ads. Therefore, views can be adopted as a valid measure over time of ad exposure.

Based on previous reasoning on recall, likeability and number of views, and their relationship with neurophysiological tools, therefore:

– H1: Eye tracking, HRV and brain activity capture (a) recalled and non-recalled online ads, (b) liked and disliked online ads, and (c) ads with high numbers vs. low numbers of YouTube views.

1. Online polls (e.g., Ace Metrix, YouGov) are becoming popular in the industry and have been used for academic purposes. Established scores in the market are considered by marketers as a valuable source of information. Ace Metrix is a company that tests ads within 24–48 h of each ad's initial airing and provides meaningful scores. Ace Metrix component scores range between 1 and 950. Scores for the general population are normally distributed and approximately centered around 530. For obtaining Ace Score, every ad is shown to a unique set of 500 respondents who complete standardized surveys that assess the ad both quantitatively and qualitatively providing rich insights by demographic segment. All ads are scored on 6 factors most likely to influence consumer behavior including relevance, likeability, information, change, attention and desire—as well as re-watch ability, purchase intent and brand linkage. After respondent scores are collected, Ace Metrix component scores are computed and assigned to each ad, creating an overall score called the Ace Score. Since its inception, Ace Metrix has consistently used the same methodology to measure the effectiveness of every ad they have tested. As a result, they are in a unique position in the field of advertising effectiveness to assess relative advertising performance between any competitive set of ads imaginable, both across different industries and different time periods.

2. In addition, the digital world is changing data availability and triggering new analytical research methods where new research avenues for unstructured data, including neurodata, and new methods like ANN call for new research (Wedel and Kannan, 2016). Venkatraman et al. (2015) conducted a similar study but in the offline context. They found that deceleration correlated with liking ($r = 0.37$, $p < -05$) and recognition ($r = 0.34$, $p < 0.05$). The current study differs in two ways: (i) we parse neurophysiological correlations in a different context,

such as digital exposure featured by searching rather and displaying; (ii) Venkatraman et al. (2015) captured advertising effort, through GRPs, and advertising outcome, through advertising elasticities; however, our study attempts to find out the correlates of neurophysiological metrics from ET, HRV and EEG with independent variables of ad effectiveness based on both, survey data and digital views. Therefore, we predicted:

– H2: Eye tracking, HRV and brain activity capture correlate with (a) self-reported score of ad effectiveness and (b) online views on YouTube.

As stated before, YouTube as a digital channel to watch commercials has almost been neglected with some related exceptions (Verhellen et al., 2013). Our aim here is to use ANN to predict the number of online views of ads placed in YouTube, where the input variables are metrics from ET, HRV, and EEG. ANN is useful for parsing non-linear relationships and adopt feed forward and back propagation approaches (West et al., 1997). ANN has been successfully applied in advertising since the mid-nineties (Curry and Moutinho, 1993). Research posits the superiority of these methods over other statistical approaches. Surprisingly, this is the first attempt to use ANN to predict online views. In our study, we aim to classify and to predict online views based on metrics from ET, HRV, and EEG. Therefore, we predicted:

– H3: Artificial neural networks based on mixed data from ET, HRV and brain activity predict the number of online views on YouTube.

MATERIALS AND METHODS

Participants and Design

Final sample consisted of 35 randomly healthy volunteers (15 women and 20 men, mean age = 25; SD = 5 years) recruited from the city where the lab is located. Initial sample measured was 47 subjects but after an examination of the dataset was carried out, 12 participants were removed due to corrupted data from experimental sessions in some of the acquired signals. All of the participants showed corrected-to-normal vision and hearing. They were asked to pay attention to the documentary as in a common situation. No mention of the importance of the ads was made. The study was approved by the Institutional Review Board of the *Polytechnic University of Valencia* with written informed consent from all subjects in accordance with the Declaration of Helsinki.

The experiment was conducted in a neuromarketing lab of a large European university and comprises the three parts shown in **Figure 1**. In Parts 1 and 2, participants sat comfortably on a reclining chair with a 32-channel EEG device, with two electrodes to measure heart variability and an eye-tracker (**Figure 1**). In Part one, participants were exposed to a mindfulness audio designed by experts to help them relax and disconnect from past experiences of the day (Fjorback et al., 2011; Demarzo et al., 2014). Then, in Part two they were shown a 30-min long

FIGURE 1 | Participant in the study. (**Top**: The EEG cap is visible. ECG electrodes placed on the chest and TMSI equipment). (**Bottom**: Eye tracking equipment is shown).

FIGURE 2 | Stimuli and experimental design.
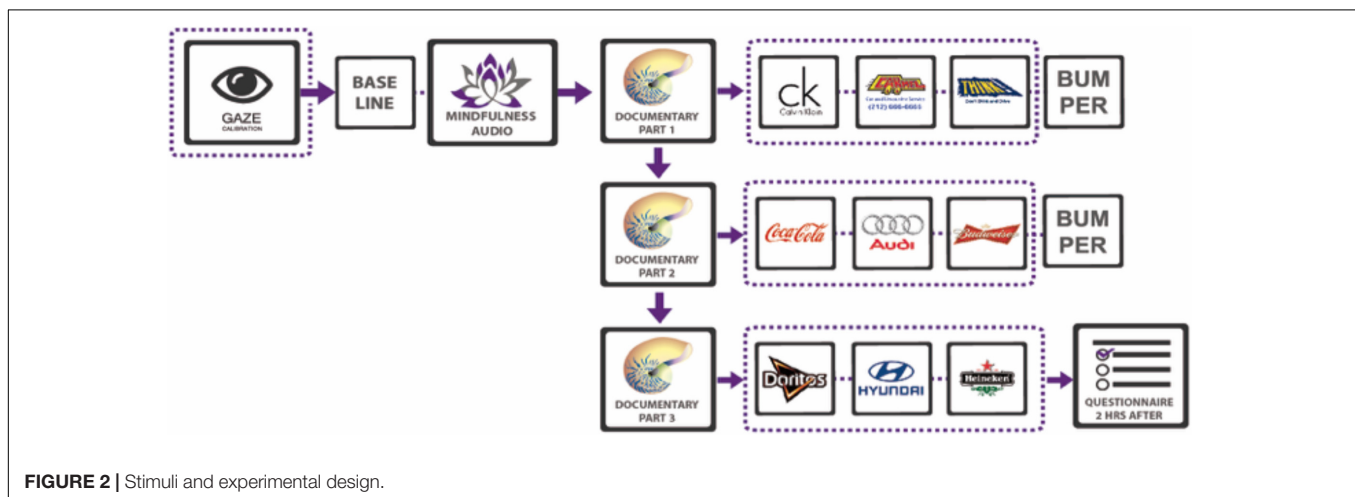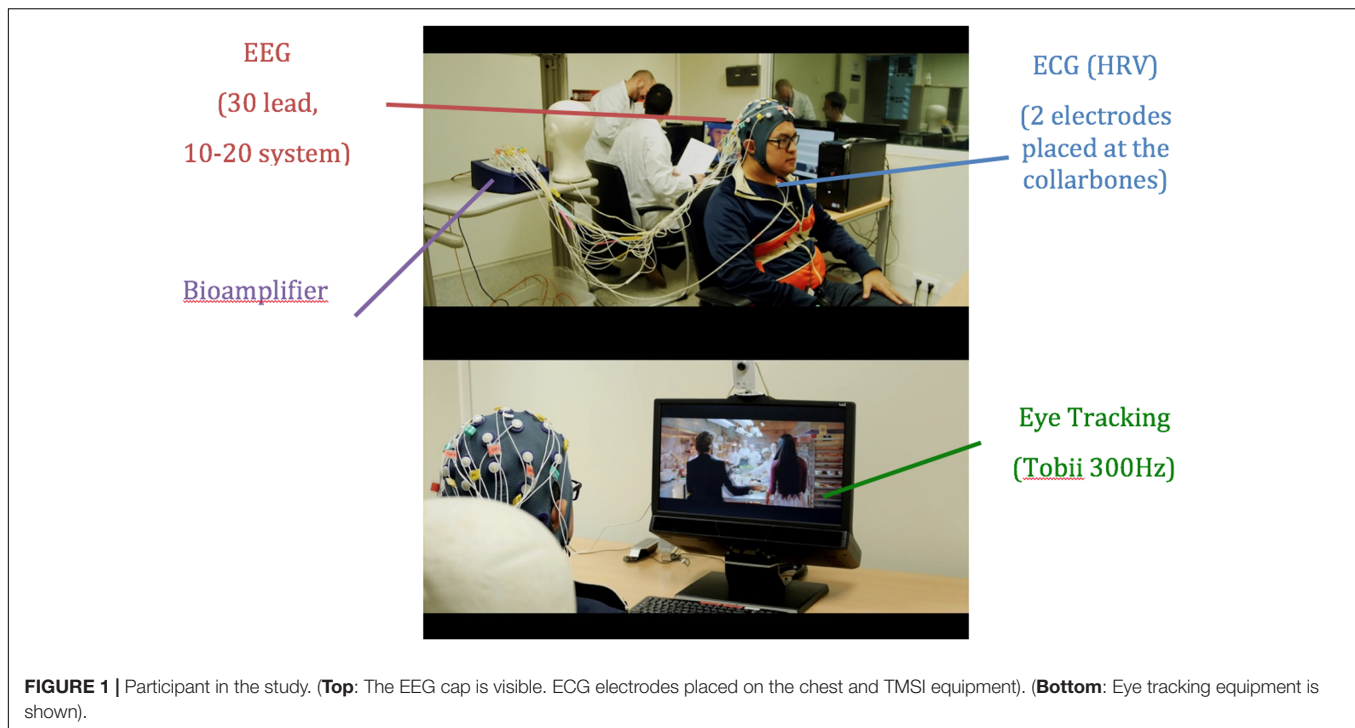
documentary with three commercial breaks of three ads lasting about 30 s each; the first break occurred after 7 min, the second in the middle of the documentary, and the third 7 min before the end as **Figure 2** depicts. At the end of this second part, participants were informed that an interview would be held 2 h later (Part 3).

Pretesting and Stimuli

Television commercials from the final of the Super Bowl 2015 were chosen because they are a good representation of the most searched high impact ads. We wait for a year to get results of number of views on the brand's official YouTube channel and a final selection of eight from 47 ads was made to represent a uniform distribution of ads ranked by number of views and

also to represent a distribution of different commercial products (**Table 1**). In addition to the number of views, the ACE Metrix[1] score was also obtained for some ads. The ACE Metrix is the most employed US scale drawn up by consumers that evaluates the ad creative effectiveness based on viewer's reaction to national TV ads. The results are presented on a scale of 1–950. The selected television commercials belong to international brands of commercial products such as drinks (3), food (1), cars (2), textiles (1) and services (1). None of the ads had been broadcasted in the country where the experiment was performed in order to remove previously unchecked exposure of the subjects to the proposed stimuli. These videos were randomly distributed

---

[1]http://www.acemetrix.com/

during the sessions with participants to avoid bias in data analysis.

Data Recording and Processing
Cerebral Recording (EEG)
Electrical activity of the brain was recorded by a stationary 32-channel system (REFA 32, TMSI hardware). EEG activity was gathered at a sampling rate of 256 Hz. The experiment used 30 Ag/AgCl water-based electrodes and bracelets attached to the opposite wrist of the subjects' dominant hand. The montage of brain electrodes followed the international 10–20 system (Jasper, 1958).

The EEG baseline was removed and channels detected as having corrupted data were rejected and interpolated from the closest electrodes (Colomer Granero et al., 2016). When a channel with erroneous data was identified, kurtosis was employed computing the fourth standardized moment in the signal of each electrode. This kurtosis is defined in Equation 1

$$K(x) = \frac{\mu_4}{\sigma_4} = \frac{E[(x-\mu)^4]}{E[(x-\mu)^2]^2} \tag{1}$$

where μ_4 was the fourth moment of the mean, σ was the standard deviation and E[x] was the expected value of signal x. The EEG signal was segmented in one second acquired accordingly to the experiment events. The intra-channel kurtosis level of each epoch was used to reject the epochs with high levels of noise.

To detect artifacts from eye movements, blinking and muscular activation, Independent Component Analysis (ICA) (Gao et al., 2010) and automatic method (ADJUST) (Mognon et al., 2011) were implemented. Each EEG artifact-free trace was band pass filtered twice in order to isolate only the spectral components in delta (1–3 Hz), theta (4–7 Hz), alpha (8–12 Hz), beta (13–24 Hz), beta extended (13–40 Hz) and gamma bands (25–60 Hz).

To quantify the cerebral activity in each band, the Global Field Power (GFP) (Wackermann et al., 1993) was calculated as explained in previous work (Colomer Granero et al., 2016).

Recent studies have shown that the main areas involved in the phenomena of memorization and pleasantness are the frontal areas (Astolfi et al., 2009). For that reason, calculation

TABLE 1 | ACE metrix score, number of visits on a brand's official YouTube channel during a year and ranking of visits established for the selected video advertisements regarding number of visits.

	ACE_score	No. visits	Rank of visits*
Drink 2	665	12,037,414	D
Car 2	394	10,583,315	D
Drink 3	Non-evaluated	10,356,789	D
Drink 1	641	6,652,069	C
Food	626	4,594,507	B
Textile	362	2,952,014	B
Car 1	611	1,453,022	B
Service	167	15,340	A

*VISIT_RANK: A: = <1M views; B = 1M–5M views; C = 5M–10M views; D = >10M views.

of electrodes in frontal lobe were taken into account. A GFP signal was then calculated for each frequency band considered in the experiment. GFP associated with each ad analyzed and GFP during a period taken as the baseline of watching a 2-min neutral documentary before the block of ads were compared and normalized to obtain the corresponding *z-score* index.

In addition to the *z-score* of GFP for each EEG band, two metrics applied in advertising research were also calculated: The *Pleasantness Index (PI)*, and the *Interest Index (II)*. The PI is a metric calculated over time that provides information about the pleasantness of the stimuli presented (Vecchiato et al., 2011). Brain activity gathered by left-frontal electrodes is compared with brain activity registered by the right-frontal electrodes (frontal asymmetry). These comparisons are made with GFP in the theta and alpha bands, comparing asymmetric pairs of electrodes.

The questionnaire generated the likeability score for each ad under study. Using this information, participants were segmented into two groups: "LIKE" and "DISLIKE." Then, the brain Pleasantness Index, PI was calculated for each group as describes (Colomer Granero et al., 2016).

The II, enables an advertising evaluation of user interest in theta and beta bands (Vecchiato et al., 2010). The most relevant peaks of these signals are selected. Two parameters were obtained: the number of peaks during a particular ad (PN_{total}) and the number of peaks during the periods the brand name appeared in that particular ad (PN_{brand}). Accordingly, the II was calculated as describes (Colomer Granero et al., 2016).

Heart Rate Variability
To analyze HRV, the electrocardiogram, ECG signal needs to be filtered (Blanco-Velasco et al., 2008), analyzed to detect QRS zones (Pan and Tompkins, 1985) and revised manually by an expert, because the appearance of a single ectopic can produce variations in certain key parameters extracted from this analysis (Clifford, 2006). HRV analysis can generate a set of metrics that can be extracted from different dimensions: time, frequency, time-frequency and non-linear. Parameters extracted from the time domain used in this study were: average heart rate (t_meanHR), standard deviation of continuous HR values (t_sdHR), the square root of the sum of successive differences between adjacent RR intervals (t_RMSSD) and the number of successive pairs of RR intervals showing a difference of more than 50 ms between them (t_NNx).

Power spectral density (PSD) analysis provides information about the amount of power in the frequency bands defined for the beat-to-beat interval signal generated. In this case, we employed the Lomb-Scargley method (Castiglioni and Di Rienzo, 1996). The frequency bands defined (low frequency, LF and high frequency, HF) are stated by Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). Power metrics can be presented in absolute values (aLF, aHF, aTotal), normalized to total energy (nLF, nHF) or in a percentage value of total energy (pLF, pHF). The ratio established between the LF and HF band provided information on the sympathetic/parasympathetic balance. The power value of the peak in the fundamental frequency (peakLF, peakHF) was also extracted.

Non-linear analysis was run using techniques such as the Poincaré graph that give us SD1 and SD2 metrics (Fishman et al., 2012). Sample entropy (sampen) is another non-linear technique that attempts to quantify the complexity or degree of new information generated (Richman and Moorman, 2000). If entropy is equal to zero, then consecutive sequences are identical. Similarly, bigger values show higher complexity of the analyzed signal.

To summarize, for each ad analyzed by subject, HRV metrics were computed by means of a computational analysis plug-in based on Matlab (Guixeres et al., 2014). For the purpose of this study, the seventeen most relevant metrics employed in HRV analysis (Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996) were selected based on time, frequency and non-linear domains.

Eye Tracking

The Tobii TX300 eye tracker[2] was used in this experiment as **Figure 1** depicts. This eye-tracker collects gaze data at 300 Hz. The subsequent analysis of raw data used Tobii Studio 3.2 software. For each commercial the following metrics were obtained from the gaze data: (i) number of fixations during each ad (Fix_Count_Advert); (ii) average duration of fixations during an ad (Fix_Dur_Advert); Furthermore, we obtained several metrics from the times the brand appeared in each ad. To calculate such metrics, a dynamic Area of Interest (AOI) that followed the brand was created using TOBII studio software to obtain (iii) the average duration of fixations exclusively focused on the brand (Fix_Dur_Br); (iv) the number of fixations during the brand appearance that focused on it (Fix_Count_Br); (v) the time from the appearance of the brand until it was fixated on for the first time (FFIX_Dur_Br); (vi) the number of visits inside the brand's AOI (Visit_Count_Br); and (vii) average duration of visits to the brand's AOI (Visit_Dur_Br). For the last two metrics, it should be remembered that a visit is the event that starts when the eye enters an AOI until it leaves such AOI. In addition to these metrics, two new metrics are proposed in this study: (viii) Number of Quadrants per second (Quad_sec); and (ix) the Gaze Brand Effectiveness Ratio (Brand_ratio).

Quad_sec enables the way the user explores a space with his eyes to be quantified. To calculate this metric, the screen surface was divided into a grid of 4×4 equal-sized quadrants. Then, the average quantity of different quadrants that the eye visited per second was calculated for each commercial. Higher values for this metric meant that the subject explored the space in "ambient mode," covering all the space with his eyes, whilst lower values meant that the subject explored the space in "focus mode," centering his visual attention on specific zones. These two modes of watching an image stimulus have been reported in previous works (Bradley et al., 2011; Holmqvist et al., 2011).

$$\text{Quad}_{seg} = \frac{N_q}{t_s} \qquad (2)$$

----
[2]Tobii.com

where N_q is the number of visits to quadrants during stimulus presentation and t_s is the duration in seconds of the stimulus.

Brand_ratio enables the effectiveness of visual attention toward the brand during the ad to be quantified. To calculate Brand_ratio, brand appearance was controlled by setting an AOI around the brand every time it appeared in the commercial. Then this metric was defined as the number of seconds that the subject looked directly at the brand divided by the total time that the brand was present on the screen during the commercial. This metric could be related to the participant's interest in and familiarity with the brand as it relates to the time that eye and brain are able to identify a brand, a concept related to familiarity (Kent and Allen, 1994).

$$\text{Brand}_{ratio} = \frac{t_{bf}}{t_b} \qquad (3)$$

where t_{bf} is the time in seconds that the gaze fixed on the brand and t_b is the total time in seconds that the brand appeared during the ad.

Questionnaire

Data were sorted using three criteria: (i) spontaneous ad recall; (ii) ad liking; (iii) the number of online views was used as a control variable. The criterion for ad recall was to remember without clues brand names of the commercials 2 h after the study. Accordingly, participants were split into two subgroups. The first dataset was related to the biometric activity collected during the observation of the recalled commercials 2 h after being exposed to them. This dataset was named RMB. The second subset included the biometric activity collected during observation of the non-recalled commercials, (FRG). The ad liking criterion was related to the biometric activity collected during observation of the television commercials that the subjects rated 5 or above on a 10-point Likert scale, being this subset named LIKE and DISLIKE, respectively.

Statistical Methods

In order to test comparisons of means of the metrics calculated, a set of Shapiro-Wilk tests (W) were conducted to test whether dependent variables deviated from normality. Then a statistical analysis was carried out using ANOVA for metrics with normal distribution and the Mann Whitney non-parametric test for metrics that did not show a normal distribution. A corrected p-value less than $p = 0.005$ was chosen to correct multiple comparisons effect (Feise, 2002).

To get the correlation of neurometrics with the ACE score and the number of online views of the ads on YouTube, a Pearson correlation was applied to the number of visits so it could be considered as a linear scale. However, Spearman's correlation was applied to the ACE score so it could be considered as a rank variable instead of a linear scale.

Artificial Neural Networks

We adjusted two neural networks with SPSS statistics using all metrics defined for EEG, HRV and ET, and including gender, as our input variables. The first network was adjusted to classify advertising responses into a ranking for the number of views

on YouTube (RANK_VISITS). This ranking divided ads into four clusters (<1M: ads with less than 1 million visits, 1M–5M, 5M–10M, and >10M: ads with more than 10 million visits). The second network was adjusted to predict the real number of visits for each ad.

Two kinds of network were compared with the same data. Multi-layer Perceptron networks (MLP) and Radial Basis Function networks (RBF). Regarding accuracy for classification and estimation of the data, MLP networks were selected finally for the two purposes instead of RBF. After testing results changing several parameters in MLP architecture, a final structure was chosen for the both neural networks (see Appendix). To validate accuracy of networks, cross-validation technique was employed. Entire sample was divided into two groups (70% of cases for training the network and 30% of cases to assess classification accuracy) and that validation was repeated 10 times ($k = 10$), selecting each time different groups for training and assessing. Final results were averaged from the 10 turns.

RESULTS

Biometric Mean Comparison

In order to test H1, we conducted a comparison among means of the metrics calculated from EEG, HRV and ET. As stated earlier, these metrics were compared by means of the following two conditions, recall vs. non-recall after 2 h (RMB vs. FRG) and by likeability (LIKE vs. DISLIKE).

Brain Response Comparison

Figure 3 shows the comparison among means for *z*-score indexes in each frequency band for the different factors chosen. In the case of remembered ads, the RMB group shows significant differences compared to the FRG group with higher values in the delta, theta, beta ext. and gamma bands. The LIKE group showed significant differences compared to the DISLIKE group with higher values in the delta, theta, beta ext. and gamma bands.

In the pleasant index (PI) and II, there were no significant differences in the comparison between the recalled and liked groups.

HRV Response Comparison

Table 2 shows the comparison of means for HRV metrics for the different factors chosen. In the case of the remembered ads, the RMB group showed significant differences compared to the FRG group, with higher values in the non-linear SD2 Poincaré index (p_SD2) that reflect higher continuous beat-to-beat variability (Piskorski and Guzik, 2007). The LIKE group showed significant differences compared to the DISLIKE group with higher values in the energy of the low frequency band (f_aLF_lomb) which is associated with sympathetic activation (Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996).

ET Response Comparison

Table 3 shows the comparison of means for eye-tracking metrics for the different classic metrics and the two-new metrics

proposed. In the case of the recalled ads, the RMB group showed significant differences vs. the FRG group, generating lower values in (Visit_Dur_Br). The LIKE group did not show significant differences compared to the DISLIKE group.

Hypothesis 1 is confirmed, as there are significant differences in each signal (EEG, HRV and ET) between ads (i) recalled and non-recalled (22% of comparisons) and (ii) liked and disliked (16% of comparisons).

As regards H2, centered on the correlation of neurometrics with the ACE score and the number of online views of the ads on YouTube, 19 of the 25 metrics showed significant ($p < 0.01$) correlation with the ACE score and 15 of the 25-metrics showed significant correlation with the number of online views. In EEG, the *z*-score in the delta band correlated with both indexes and the *z*-score in the theta band correlated with the ACE score. Both indexes, pleasantness and interest, showed high values in terms of significant correlation with the ACE score. In particular, there was a high level of correlation between PI_theta and the number of visits. In terms of ET, the proposed metric (Brand_ratio) showed significant correlations with both indexes, PI and II. In addition, Quad_sec showed significant correlation with the ACE_score. Visit duration and Fixation count on the brand showed significant correlation with the number of visits. Fixation duration showed a negative correlation with the ACE score and the value of the correlations of the Fixation count during the ad for both indexes was relevant. Focusing on HRV, t_NN50 showed significant correlations with both indexes. The total energy band; LF and HF bands, also showed correlations with both indexes. The normalized values of LF and HF and the sympathovagal index (LFHF) showed significant correlations with the ACE score. Frequency with the maximum peak on the HF band (peakHF) and the SD2 Poincaré index also showed significant correlations with the number of visits. The type 1 non-linear parameter sample entropy index (sampen1) showed significant correlations with both indexes whilst sampen2 only correlated with number of visits. Therefore, H2 is confirmed, showing significant correlations between metrics for EEG, HRV and ET with ACE score and number of visits on YouTube.

Predicting Ad Effectiveness on the Internet with ANN

ANN to Rank Visit Classification

Two kinds of ANNs were compared with the same data: MLP and RBF. In terms of accuracy when classifying data, MLP networks worked better than RBF. After testing results and changing several parameters in the MLP architecture, the final structure chosen had one hidden layer with a tangent hyperbolic activation function and an output layer with a softmax activation function. All the co-variables were typified (see Appendix for more details).

Table 4 shows the final results for the classification of the training and the test dataset using the neural network. The percentage of correct predictions in the test dataset was 82.9%. Three higher ranking levels (1M–5M, 5M–10M, and >10M) showed high effectiveness ratios of classification (88.5, 100, and 90%, respectively). Only the first group (<1M) showed a poor

FIGURE 3 | Results from brain response (EEG) comparing GFP results for 5 specific bands (*values with significant differences $p < 0.005$; NP, non-parametric test carried out).
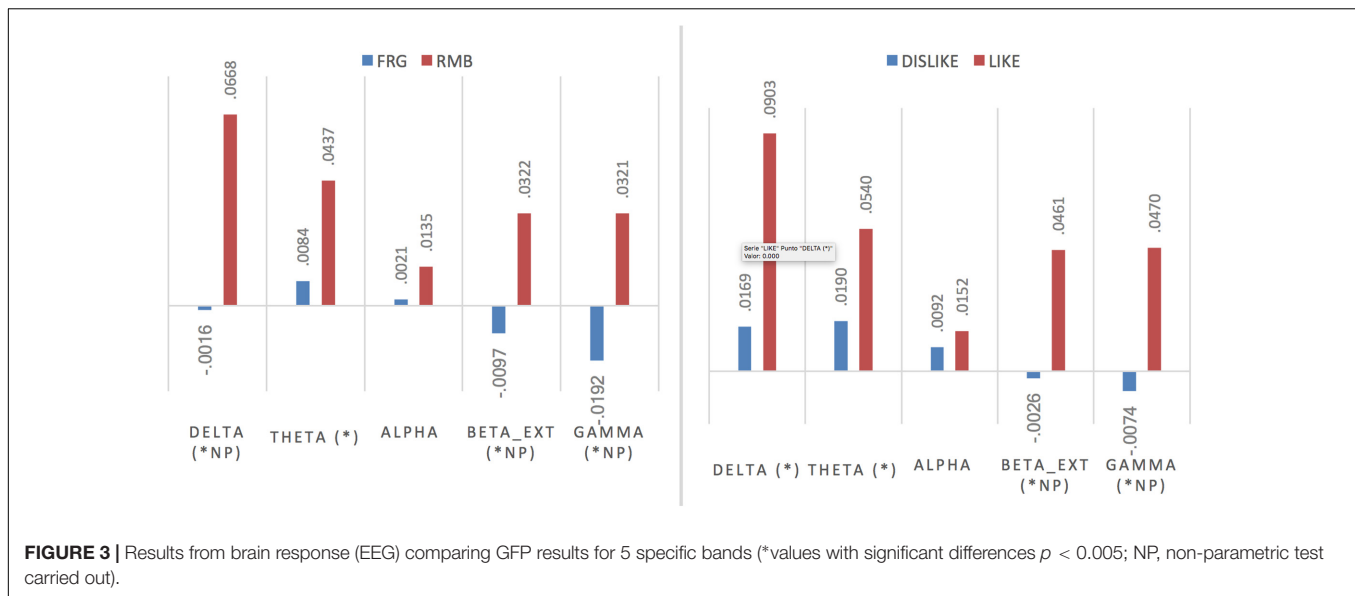
effectiveness ratio of 36.4%, mistaking more than 54% of cases for second level (1M–5M).

Of the 37 input variables used in neural networks, the importance of each normalized metric to predict was extracted from the neural network. This metric shows the most relevant variables for classifying each case according to the correct ranking. The pleasantness index extracted from the theta band (PI_theta, 100%) was the most important parameter for brain metrics, followed by II_theta (56.6%) and PI_ alpha (56.5%). Regarding HRV metrics, the mean of HR (64.20%) was the most representative parameter, followed by SD1 Poincaré (61.40%) and

t_RMSSD (58.30%). In terms of eye-tracking metrics, the fixation count during the ads (59.30%) was the most important index, followed by the average fixation count during brand appearance (59.30%), and the average duration of fixations during brand appearance (45.30%). Gender inclusion and whether the ad was consciously remembered came last in the ranking with very little importance. The proposed metrics from ET showed medium importance (Ratio_Brand: 40.80% and Quad_sec: 40.20%). When comparing the frequency brain bands, the theta band behaved best, in the PI index (100%), II index (56.6%) and z-score (37.50%).

TABLE 2 | Mean comparisons of HRV metrics in the time, frequency and non-linear domains.

HRV Metrics	RMB		FRG			UNLIKE		LIKE		
	Mean	SD	Mean	SD	p	Mean	SD	Mean	SD	p
t_meanHR	71.9	12.3	70.6	8		69.6	8.8	71.1	7.8	
t_sdHR	5.2	7.3	5.5	4.5		5.1	5.6	5.3	2.7	
t_NN50	13.3	18.4	13.8	12.7		13	13.6	13.5	13.3	
t_RMSSD	43.5	32.3	54.6	112.3		43.2	26	60.1	142.7	
f_aLF_lomb	0.031	0.015	0.033	0.013		0.031	0.014	0.035	0.014	(*NP)
f_aHF_lomb	0.027	0.020	0.027	0.017		0.029	0.021	0.026	0.017	
f_aTotal_lomb	0.060	0.027	0.061	0.022		0.062	0.028	0.062	0.024	
f_nLF_lomb	0.539	0.223	0.573	0.2		0.547	0.203	0.594	0.194	
f_nHF_lomb	0.461	0.223	0.427	0.2		0.453	0.203	0.406	0.194	
f_LFHF_lomb	2.022	2.156	2.121	2.057		1.786	1.545	2.412	2.412	
f_peakLF_lomb	0.1	0.02	0.09	0.02		0.09	0.02	0.09	0.02	
f_peakHF_lomb	0.27	0.07	0.26	0.07		0.26	0.07	0.25	0.07	
p_SD1	30.7	22.8	32.8	21.4		30	18	32.9	19.8	
p_SD2	66.1	28	77.6	36.7	(*NP)	68.8	30.4	80.3	38	
p_SD1vsSD2	0.497	0.382	0.441	0.192		0.462	0.224	0.423	0.154	
nl_sampen1	2.175	0.227	2.141	0.243		2.179	0.25	2.143	0.217	
nl_sampen2	1.803	0.486	4.19	0.532	(*NP)	9.373	0.788	1.66	0.389	

*Values with significant differences $p < 0.005$; NP, non-parametric test.

ANN for Estimating the Number of Online Views

As described above, MLP and RBF were compared only in terms of accuracy (time execution was extremely low in both cases). MLP networks also worked better than RBF networks in terms of accuracy when estimating the number of visits. After testing results and changing several parameters in the MLP architecture, the final structure chosen had one hidden layer with a tangent hyperbolic activation function and an output layer with an identity activation function. All the co-variables and the dependent variable were typified.

For the final MLP structure, the first subset was used to train the MLP and the estimated values of the number of visits were obtained from the second subset to test the accuracy of the network. Final results for the estimation of training and the test dataset using an ANN. Relative error from the test dataset was 0.199, that is, a significant level of variance.

Of the 37 input variables used in the neural networks, the pleasantness index extracted from the theta band (PI_Theta, 100%) was again the most relevant variable in the neural network in predicting the number of views. Regarding brain metrics, PI_theta was followed by PI_alpha (13.10%) and II_theta (11.60%), though these were a long way behind. In HRV metrics, sample entropy (53.70%) was the most important parameter, followed by the total energy of frequency band (43.40%) and t_RMSSD (41.70%). Regarding eye-tracking metrics, the count of fixations during ads (44.30%) was the most important index, as on the first occasion, followed by number of visits during brand appearance (27%) and then followed very closely by the count of fixations during brand appearance (26.60%). The inclusion of gender in the model and whether the ad was consciously recalled ranked last and do not improve prediction. The proposed new metrics extracted from ET showed different results. Quad_sec showed a medium-low importance (16%) and Ratio_Brand a low importance (6.70%). Half the HRV metrics were in the upper positions.

Figure 4 represents predicted vs. observed values as a scatterplot for each case. Figure shows that neural network classifier worked well in classifying real data regarding biometric response, specifically observing the cluster of ads with higher number of views vs. ads with poor audience. There was some dispersion in all the ads in terms of real values.

TABLE 3 | Mean comparisons of ET metrics.

Eye Tracking Metrics	RMB		FRG			UNLIKE		LIKE		
	Mean	SD	Mean	SD	p	Mean	SD	Mean	SD	p
Brand_ratio	0.36	0.25	0.32	0.25		0.37	0.26	0.31	0.24	
Visit_Count_Br	6.7	2.42	6.65	2.33		7.29	2.53	6.32	2.38	
Visit_Dur_Br	0.76	0.24	0.63	0.25	(*NP)	0.75	0.23	0.62	0.38	
FFix_Dur_Br	0.18	0.14	0.16	0.13		0.17	0.13	0.15	0.13	
Fix_Dur_Br	0.18	0.12	0.17	0.11		0.18	0.15	0.17	0.12	
Fix_Count_Br	28.23	28.97	23.73	22.87		33.76	37.48	21.98	19.29	
Quad_sec	2.47	1.35	2.48	1.29		2.65	1.62	2.58	1.24	
Fix_Dur_Advert	0.24	0.1	0.24	0.11		0.24	0.12	0.21	0.11	
Fix_Count_Advert	273.32	120.34	281.8	200.12		340.73	180.73	270.83	165.43	

*Values with significant differences p < 0.005; NP, non-parametric test.

TABLE 4 | Results for the successful classification of training and test dataset applying an adjusted neural network.

		Classification				
	Real rank order	Predicted				
		<1M	1M–5M	5M–10M	>10M	% succ.
Training	<1M	17	1	0	0	94.4%
	1M–5M	0	60	0	0	100.0%
	5M–10M	0	0	23	1	95.8%
	>10M	0	0	0	67	100.0%
	Overall percentage	10.1%	36.1%	13.6%	40.2%	98.8%
Test	<1M	4	6	1	0	36.4%
	1M–5M	2	23	0	1	88.5%
	5M–10M	0	0	9	0	100.0%
	>10M	0	0	3	27	90.0%
	Overall percentage	7.9%	38.2%	17.1%	36.8%	82.9%

Number of cases in each possible class and percentage of successful cases.
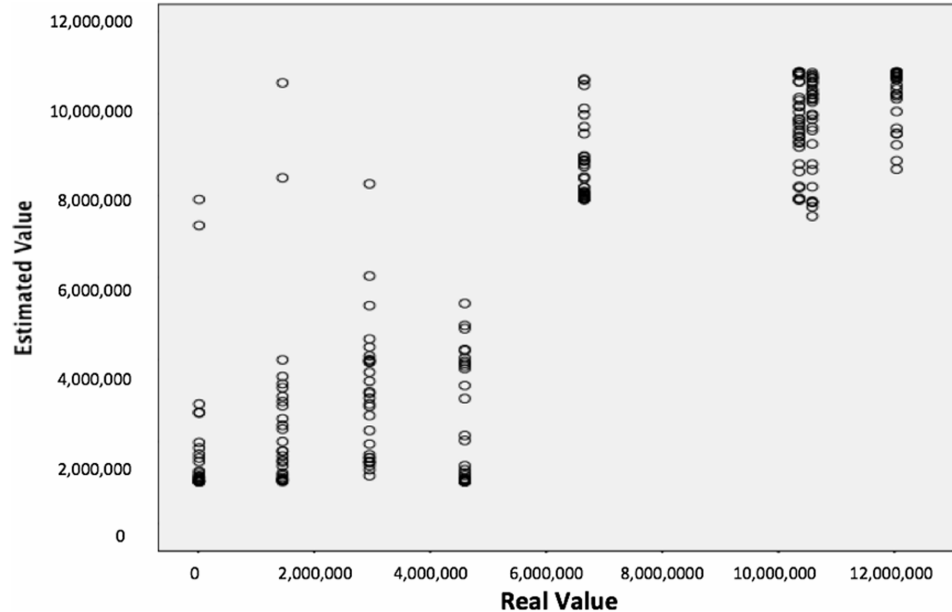
FIGURE 4 | Scatterplot showing predicted vs. observed-values for number of visits.

DISCUSSION

In this study, several hypotheses have been confirmed. The first hypothesis was supported, showing significant differences in physiological responses (EEG and HRV) and ET for each of the three dimensions raised (i.e., recall, liking and visits). In cases where participants remembered the ad after 2 h, results showed higher probability that the spectral amplitude in the RMB condition was always higher than the power spectra in the FRG conditions (Astolfi et al., 2008). A statistical increase of PSD in the prefrontal and parietal areas for the RMB dataset compared with the FRG dataset was in line with the suggested role of these regions during the transfer of sensory perceptions from short-term memory to long-term memory storage. Specifically, there were higher values in the theta band for the cases where the ad was remembered, which is in line with other studies (Werkle-Bergner et al., 2006; Boksem and Smidts, 2014; Vecchiato et al., 2014). Regarding this condition, an interesting future analysis should look at the biometric response differences between people that remember the ad without remembering the brand or vice versa.

In terms of the HRV analysis, the means of the sample entropy parameters were significant, showing the higher complexity of heart variability in cases where the ad was remembered. Valenza et al. (2012), showed that Approximate Entropy, decreased during arousal elicitation using images from International Affective Picture Session (IAPS) but there are no studies until now that have related HRV entropy with remember cognitive function.

In ET, cases where ads were unrecalled showed longer duration of visits to the brand when it appeared on screen. Higher values for this metric are related to difficulty in identifying

an object (Goldberg et al., 2002; Wedel, 2013). This fact could be explained by poor identification of the brand, which could indicate that the ad will be forgotten in the short term.

Regarding cases where participants rated the ad positively or negatively, brain activity was stronger in terms of PSD in the LIKE group than in the DISLIKE group. These results are congruent with another EEG studio based on the observation of pictures from the international affective picture system (Aftanas et al., 2004). In HRV, there were only significant differences in energy in the LF frequency band associated with sympathetic activation. In ET, no significant differences were found for the proposed metrics.

The second hypothesis was supported showing significant correlations between physiological and eye-tracking responses with the ACE score and with the number of online visits. All the indexes (pleasantness and interest) calculated showed significant correlations. The pleasantness index in the theta band presented an especially high correlation with the number of visits and the ACE score. In ET, there were relevant correlations with both metrics proposed in this work. Brand_ratio showed a positive linear relation between the percentage of time watching a brand and the number of views on the Internet. The number of fixations for the brand showed a positive relation with the number of views on the Internet. Regarding HRV, there was a positive correlation with tNN50 for both outputs. All the energy values in the frequency bands (total, LF and HF) were negatively correlated with both outputs. The ACE score was positively related to the normalized LF band associated with sympathetic activation. Non-linear entropy parameters also showed a negative correlation with both outputs revealing that an increase in the complexity of the HRV signal is associated with less quality and effectiveness of the ad on the Internet.

Hypothesis three aimed to test whether ANN with relevant biometrics could represent an interesting technique to classify ads based on their ranking on the Internet and to estimate the number of visits on the Internet. The results obtained showed that the ANN were able to accurately classify and estimate the effectiveness of each ad on the Internet via their biometric response. The results for the first network, which were adjusted to classify each ad based on a four-level ranking, showed a global average accuracy of 82.9%. Poor accuracy was obtained with ads with a lower number of views but this could have been improved if more ads in this ranking had been selected in the stimulus group. The relevant metrics for this classification were the pleasantness index and II in the theta band, the mean heart rate and the SD1 Poincare in HRV, the number of fixations during the ad and inside the brand for ET. Results for the second network to estimate the number of views on the YouTube for each ad showed a relative error of 0.199. The most important metrics for this estimate were the pleasantness index in the theta band, entropy in HRV and the number of fixations during advertising in ET. Despite good results to estimate the number of views, it would seem that classifying ads according to a ranking constitutes a better approach, taking into account the excellent results obtained in the first classifier.

Further research is needed, with more studies comparing new techniques for classification, such as Linear Discriminant Analysis, Marquardt Backpropagation Algorithm, and Deep Learning. In addition, new metrics and new signals extracted from biometric responses must be tested to find out which parameters are best to evaluate the effectiveness of advertising. The group of different categories of advertising must also be increased. Future studies should also focus on adjusting personalized classifiers to advertising categories (fashion, food, social, etc.), different channels (e.g., Facebook) and formats (desktop or mobile). Also, new metrics like facial gesture coding (McDuff et al., 2015) and fNIRS (Kopton and Kenning, 2014) could be mixed in new models for predicting Ad effectiveness.

CONCLUSION

This study has shown that aspects related to the impact of advertising, such as whether the ad is going to be remembered or whether it is going to be highly rated can be detected from an analysis of consumers' biometric responses during the viewing of these ads. We also found differences in the impact of advertising in terms of gender, which encourages the use of these biometric data to design advertising content that is tailored to each individual group of population. Other variables, such as age, cultural level, and even personality could be explored in future studies to test whether there are similar differences to the ones found in gender. The final conclusion that this study has yielded is that the effectiveness of a new ad on YouTube can be predicted using metrics extracted from EEG, HRV and ET. Up until now, there has been no evidence that biometric responses can help to classify the numbers of views on YouTube for an ad. This study has also contributed with two new metrics for ET that can be used in research on advertising. These results will help to explain the success of advertising responses showing an interesting methodology to be use by practitioners designing advertising content.

AUTHOR CONTRIBUTIONS

JG is the corresponding author. JG, EB, and MAR designed the study. JAA conducted the study. JG, EB, and JAA conducted the literature review and wrote the research summaries. ACG and VNO analyzed the EEG data, JP and JAA analyzed HRV and ET data. MAR and EB are the directors of this work. JG wrote the first draft of the manuscript, and all authors contributed to and have approved the final manuscript. Authors had full access to the study data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01808/full#supplementary-material

REFERENCES

Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., and Suri, J. S. (2006). Heart rate variability: a review. *Med. Biol. Eng. Comput.* 44, 1031–1051. doi: 10.1007/s11517-006-0119-0

Aftanas, L. I., Reva, N. V., Varlamov, A. A., Pavlov, S. V., and Makhnev, V. P. (2004). Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. *Neurosci. Behav. Physiol.* 34, 859–867. doi: 10.1023/B:NEAB.0000038139.39812.eb

Astolfi, L., Fallani, F. D. V., Cincotti, F., Mattia, D., Bianchi, L., Marciani, M. G., et al. (2008). Neural basis for brain responses to TV commercials: a high-resolution EEG study. *IEEE Trans. Neural Syst. Rehabil. Eng.* 16, 522–531. doi: 10.1109/TNSRE.2008.2009784

Astolfi, L., Fallani, F. D. V., Cincotti, F., Mattia, D., Bianchi, L., Marciani, M. G., et al. (2009). Brain activity during the memorization of visual scenes from TV commercials: an application of high resolution EEG and steady state somatosensory evoked potentials technologies. *J. Physiol.* 103, 333–341. doi: 10.1016/j.jphysparis.2009.07.002

Baack, D. W., Wilson, R. T., and Till, B. D. (2008). Creativity and memory effects: recall, recognition, and an exploration of nontraditional media. *J. Advert.* 37, 85–94. doi: 10.2753/JOA0091-3367370407

Bellman, S., Murphy, J., Treleaven-Hassard, S., O'Farrell, J., Qiu, L., and Varan, D. (2013). Using internet behavior to deliver relevant television commercials. *J. Interact. Mark.* 27, 130–140. doi: 10.1016/j.intmar.2012.12.001

Bigné, E. (2016). Frontiers in research in business: will you be in? *Eur. J. Manag. Bus. Econ.* 25, 89–90. doi: 10.1016/j.redeen.2016.09.001

Bigné, E., Llinares, C., and Torrecilla, C. (2016). Elapsed time on first buying triggers brand choices within a category: a virtual reality-based study. *J. Bus. Res.* 69, 1423–1427. doi: 10.1016/j.jbusres.2015.10.119

Blanco-Velasco, M., Weng, B., and Barner, K. E. (2008). ECG signal denoising and baseline wander correction based on the empirical mode decomposition. *Comput. Biol. Med.* 38, 1–13. doi: 10.1016/j.compbiomed.2007.06.003

Boksem, M. A. S., and Smidts, A. (2014). Brain responses to movie-trailers predict individual preferences for movies and their population-wide commercial success. *J. Mark. Res.* 52, 482–492. doi: 10.1509/jmr.13.0572

Bolls, P. D., Wise, K., and Bradley, S. D. (2012). "Embodied motivated cognition: a theoretical framework for studying dynamic mental processes underlying advertising exposure," in *Advertising Theory*, eds S. Rodgers and E. Thorson (New York, NY: Routledge), 105–119.

Bradley, M. M., Houbova, P., Miccoli, L., Costa, V. D., and Lang, P. J. (2011). Scan patterns when viewing natural scenes: emotion, complexity, and repetition. *Psychophysiology* 48, 1544–1553. doi: 10.1111/j.1469-8986.2011.01223.x

Castiglioni, P., and Di Rienzo, M. (1996). "On the evaluation of heart rate spectra: the lomb periodogram," in *Proceedings of the Computers in Cardiology, IEEE*, Indianapolis, IN, 505–508. doi: 10.1109/CIC.1996.542584

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., and Moon, S. (2007). "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (San Diego, CA: Association for Computing Machinery), 1–14. doi: 10.1145/1298306.1298309

Chen, L., Zhou, Y., and Chiu, D. M. (2014). "A lifetime model of online video popularity," in *Proceedings of the 23rd International Conference on Computer Communication and Networks, IEEE*, Shanghai, 1–8. doi: 10.1109/ICCCN.2014.6911774

Christoforou, C., Christou-Champi, S., Constantinidou, F., and Theodorou, M. (2015). From the eyes and the heart: a novel eye-gaze metric that predicts video preferences of a large audience. *Front. Psychol.* 6:579. doi: 10.3389/fpsyg.2015.00579

Clifford, G. D. (2006). "ECG statistics, noise, artifacts, and missing data," in *Advanced Methods and Tools for ECG Data Analysis*, eds G. Cliford, F. Azuaje, and P. E. MsSahrry (Boston, MA: Artech House), 55–99.

Colomer Granero, A., Fuentes-Hurtado, F., Naranjo, V., Guixeres, J., Ausín, J. M., and Alcañiz, M. (2016). A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents". *Front. Comput. Neurosci.* 10:74. doi: 10.3389/fncom.2016.00074

Couwenberg, L. E., Boksem, M. A., Dietvorst, R. C., Worm, L., Verbeke, W. J., and Smidts, A. (2017). Neural responses to functional and experiential ad appeals: explaining ad effectiveness. *Int. J. Res. Mark.* 34, 355–366. doi: 10.1016/j.ijresmar.2016.10.005

Curry, B., and Moutinho, L. (1993). Neural networks in marketing: modelling consumer responses to advertising stimuli. *Eur. J. Mark.* 27, 5–20. doi: 10.1108/03090569310040325

Daugherty, T., Hoffman, E., and Kennedy, K. (2016). Research in reverse: Ad testing using an inductive consumer neuroscience approach. *J. Bus. Res.* 69, 3168–3176. doi: 10.1016/j.jbusres.2015.12.005

Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. *Biol. Psychol.* 67, 219–233. doi: 10.1016/j.biopsycho.2004.03.008

Deitz, G. D., Royne, M. B., Peasley, M. C., and Huang, J. C. (2016). EEG-based measures versus panel ratings: predicting social-media based behavioral responses to super Bowl Ads. *J. Advert. Res.* 56, 217–227. doi: 10.2501/JAR-2016-030

Demarzo, M. M. P., Montero-Marin, J., Stein, P. K., Cebolla, A., Provinciale, J. G., and García-Campayo, J. (2014). Mindfulness may both moderate and mediate the effect of physical fitness on cardiovascular responses to stress: a speculative hypothesis. *Front. Physiol.* 5:105. doi: 10.3389/fphys.2014.00105

Donthu, N., Cherian, J., and Bhargava, M. (1993). Factors influencing recall of outdoor advertising. *J. Advert. Res.* 33, 64–73.

dos Santos, R. D. O. J., de Oliveira, J. H. C., Rocha, J. B., and Giraldi, J. D. M. E. (2015). Eye tracking in neuromarketing: a research agenda for marketing studies. *Int. J. Psychol. Stud.* 7, 32–42. doi: 10.5539/ijps.v7n1p32

Elsen, M., Pieters, P., and Wedel, M. (2016). Thin slice impressions: how advertising evaluation depends on exposure duration. *J. Mark. Res.* 53, 563–579. doi: 10.1509/jmr.13.0398

Feise, R. J. (2002). Do multiple outcome measures require P-value adjustment? *BMC Med. Res. Methodol.* 2:8.

Fishman, M., Jacono, F. J., Park, S., Jamasebi, R., Thungtong, A., Loparo, K. A., et al. (2012). A method for analyzing temporal patterns of variability of a time series from Poincare plots. *J. Appl. Physiol.* 113, 297–306. doi: 10.1152/japplphysiol.01377.2010

Fjorback, L. O., Arendt, M., Ørnbøl, E., Fink, P., and Walach, H. (2011). Mindfulness-based stress reduction and mindfulness-based cognitive therapy– a systematic review of randomized controlled trials. *Acta Psychiatr. Scand.* 124, 102–119. doi: 10.1111/j.1600-0447.2011.01704.x

Gao, J. F., Yang, Y., Lin, P., Wang, P., and Zheng, C. X. (2010). Automatic removal of eye-movement and blink artifacts from EEG signals. *Brain Topogr.* 23, 105–114. doi: 10.1007/s10548-009-0131-4

Geisler, F. C., Vennewald, N., Kubiak, T., and Weber, H. (2010). The impact of heart rate variability on subjective well-being is mediated by emotion regulation. *Pers. Individ. Dif.* 49, 723–728. doi: 10.1016/j.paid.2010.06.015

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., and Wichansky, A. M. (2002). "Eye tracking in web search tasks," in *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications - ETRA*, New York, NY, 51. doi: 10.1145/507072.507082

Grandjean, D., Sander, D., and Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Conscious. Cogn.* 17, 484–495. doi: 10.1016/j.concog.2008.03.019

Guerreiro, J., Rita, P., and Trigueiros, D. (2015). Attention, emotions and cause-related marketing effectiveness. *Eur. J. Mark.* 49, 1728–1750. doi: 10.1108/EJM-09-2014-0543

Guixeres, J., Redon, P., Saiz, J., Alvarez, J., Torro, M. I., Cantero, L., et al. (2014). Cardiovascular fitness in youth: association with obesity and metabolic abnormalities. *Nutr. Hosp.* 29, 1290–1297. doi: 10.3305/nh.2014.29.6.7383

Ha, L. (2008). Online advertising research in advertising journals: a review. *J. Curr. Issues Res. Advert.* 30, 31–48. doi: 10.1080/10641734.2008.10505236

Harmon-Jones, E., Gable, P. A., and Peterson, C. K. (2010). The role of asymmetric frontal cortical activity in emotion-related phenomena: a review and update. *Biol. Psychol.* 84, 451–462. doi: 10.1016/j.biopsycho.2009.08.010

Holmqvist, K., Andrà, C., Lindström, P., Arzarello, F., Ferrara, F., Robutti, O., et al. (2011). A method for quantifying focused versus overview behavior in AOI sequences. *Behav. Res. Methods* 43, 987–998. doi: 10.3758/s13428-011-0104-x

Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* 10, 371–375.

Kent, R. J., and Allen, C. T. (1994). Competitive interference effects in consumer memory for advertising: the role of brand familiarity. *J. Mark.* 58, 97–105. doi: 10.2307/1252313

Khushaba, R. N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B. E., and Townsend, C. (2013). Consumer neuroscience: assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. *Expert Syst. Appl.* 40, 3803–3812. doi: 10.1016/j.eswa.2012.12.095

Kim, K., Hayes, J. L., Avant, J. A., and Reid, L. N. (2014). Trends in advertising research: a longitudinal analysis of leading advertising, marketing, and communication journals, 1980 to 2010. *J. Advert.* 43, 296–316. doi: 10.1080/00913367.2013.857620

Kopton, I. M., and Kenning, P. (2014). Near-infrared spectroscopy (NIRS) as a new tool for neuroeconomic research. *Front. Hum. Neurosci.* 8:549. doi: 10.3389/fnhum.2014.00549

Kühn, S., Strelow, E., and Gallinat, J. (2016). Multiple "buy buttons" in the brain: forecasting chocolate sales at point-of-sale based on functional brain activation using fMRI. *Neuroimage* 136, 122–128. doi: 10.1016/j.neuroimage.2016.05.021

Lang, A., Bolls, P., Potter, R. F., and Kawahara, K. (1999). The effects of production pacing and arousing content on the information processing of television. *J. Broadcast. Electron. Media* 43, 451–475. doi: 10.1080/08838159909364504

Lee, J., and Ahn, J. H. (2012). Attention to banner ads and their effectiveness: an eye-tracking approach. *Int. J. Electron. Commerce* 17, 119–137. doi: 10.2753/JEC1086-4415170105

McAlister, L., Srinivasan, R., Jindal, N., and Cannella, A. A. (2016). Advertising effectiveness: the moderating effect of firm strategy. *J. Mark. Res.* 53, 207–224. doi: 10.1509/jmr.13.0285

McDuff, D., El Kaliouby, R., Cohn, J. F., and Picard, R. W. (2015). Predicting ad liking and purchase intent: large-scale analysis of facial responses to ads. *IEEE Trans. Affect. Comput.* 6, 223–235. doi: 10.1109/TAFFC.2014.2384198

McKinsey (2015). *Global Media Report 2015*. Available at: www.mckinsey.com/ client_service/media_and_entertainment (accessed December 6, 2016).

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48, 229–240. doi: 10.1111/j.1469-8986.2010.01061.x

Moriarty, S., Mitchell, N., and Wells, W. (2012). *Advertising & IMC Principles and Practice*, 9th Edn. Hoboken, NJ: Pearson.

Morin, C. (2011). Neuromarketing: the new science of consumer behavior. *Society* 48, 131–135. doi: 10.1109/MPUL.2012.2189166

Mould, D., Mandryk, R. L., and Li, H. (2012). Emotional response and visual attention to non-photorealistic images. *Comput. Graph.* 36, 658–672. doi: 10.1016/j.cag.2012.03.039

Ohme, R., Reykowska, D., Wiener, D., and Choromanska, A. (2010). Application of frontal EEG asymmetry to advertising research. *J. Econ. Psychol.* 31, 785–793. doi: 10.3791/55872

Pan, J., and Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* 32, 230–236. doi: 10.1109/TBME.1985.325532

Pieters, R., Warlop, L., and Wedel, M. (2002). Breaking through the clutter: benefits of advertisement originality and familiarity for brand attention and memory. *Manage. Sci.* 48, 765–781. doi: 10.1287/mnsc.48.6.765.192

Piskorski, J., and Guzik, P. (2007). Geometry of the Poincaré plot of RR intervals and its asymmetry in healthy adults. *Physiol. Meas.* 28, 287–300. doi: 10.1088/ 0967-3334/28/3/005

Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* 278, H2039–H2049.

Rossiter, J. R., and Eagleson, G. (1994). Conclusions from the ARF's copy research validity project. *J. Advert. Res.* 34, 19–33.

Ruanguttamanun, C. (2014). Neuromarketing: I put myself into a fMRI scanner and realized that I love Louis Vuitton Ads. *Procedia* 148, 211–218. doi: 10.1016/ j.sbspro.2014.07.036

Shehu, E., Bijmolt, T. H., and Clement, M. (2016). Effects of likeability dynamics on consumers' intention to share online video advertisements. *J. Interact. Mark.* 35, 27–43. doi: 10.1016/j.intmar.2016.01.001

Smith, A. N., Fischer, E., and Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *J. Interact. Mark.* 26, 102–113. doi: 10.1016/j.intmar.2012.01.002

Strach, P., Zuber, K., Fowler, E. F., Ridout, T. N., and Searles, K. (2015). In a different voice? Explaining the use of men and women as voice-over announcers in political advertising. *Polit. Commun.* 32, 183–205. doi: 10.1080/10584609. 2014.914614

Task Force of the European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Eur. Heart J.* 17, 354–381. doi: 10.1093/oxfordjournals.eurheartj.a014868

Tomkovick, C., Yelkur, R., and Christians, L. (2001). The USA's biggest marketing event keeps getting bigger: an in-depth look at Super Bowl advertising in the 1990s. *J. Mark. Commun.* 7, 89–108. doi: 10.1080/13527260121725

Vakratsas, D., and Ambler, T. (1999). How advertising works: What do we really know? *J. Mark.* 63, 26–43. doi: 10.2307/1251999

Valenza, G., Allegrini, P., Lanatà, A., and Scilingo, E. P. (2012). Dominant Lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation. *Front. Neuroeng.* 5:3. doi: 10.3389/fneng.2012. 00003

Valenza, G., Citi, L., Lanatá, A., Scilingo, E. P., and Barbieri, R. (2014). Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Sci. Rep.* 4:4998. doi: 10.1038/srep04998

Varan, D., Lang, A., Barwise, P., Weber, R., and Bellman, S. (2015). How reliable are neuromarketers' measures of advertising effectiveness? *J. Advert. Res.* 55, 176–191. doi: 10.2501/JAR-55-2-176-191

Vecchiato, G., Astolfi, L., Tabarrini, A., Salinari, S., Mattia, D., Cincotti, F., et al. (2010). EEG analysis of the brain activity during the observation of commercial, political, or public service announcements. *Comput. Intell. Neurosci.* 2010:985867. doi: 10.1155/2010/985867

Vecchiato, G., Maglione, A. G., Cherubino, P., Wasikowska, B., Wawrzyniak, A., Latuszynska, A., et al. (2014). Neurophysiological tools to investigate consumer's gender differences during the observation of TV commercials. *Comput. Math. Methods Med.* 2014:912981. doi: 10.1155/2014/912981

Vecchiato, G., Susac, A., Margeti, S., Fallani, F. D. V., Maglione, A. G., Supek, S., et al. (2013). High-resolution EEG analysis of power spectral density maps and coherence networks in a proportional reasoning task. *Brain Topogr.* 26, 303–314. doi: 10.1007/s10548-012-0259-5

Vecchiato, G., Toppi, J., Astolfi, L., Fallani, F. D. V., Cincotti, F., Mattia, D., et al. (2011). Spectral EEG frontal asymmetries correlate with the experienced pleasantness of TV commercial advertisements. *Med. Biol. Eng. Comput.* 49, 579–583. doi: 10.1007/s11517-011-0747-x

Venkatraman, V., Dimoka, A., Pavlou, P. A., Vo, K., Hampton, W., Bollinger, B., et al. (2015). Predicting advertising success beyond traditional measures: new insights from neurophysiological methods and market response modeling. *J. Mark. Res.* 52, 436–452. doi: 10.1509/jmr.13.0593

Verhellen, Y., Dens, N., and De Pelsmacker, P. (2013). Consumer responses to brands placed in youtube movies: the effect of prominence and endorser expertise. *J. Electron. Commerce Res.* 14, 287–303.

Wackermann, J., Lehmann, D., Michel, C. M., and Strik, W. K. (1993). Adaptive segmentation of spontaneous EEG map series into spatially defined microstates. *Int. J. Psychophysiol.* 14, 269–283. doi: 10.1016/0167-8760(93)90041-M

Wedel, M. (2013). *Attention Research in Marketing: A Review of Eye Tracking Studies*. 1–28. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_ id=2460289

Wedel, M., and Kannan, P. K. (2016). Marketing analytics for data-rich environments. *J. Mark.* 80, 97–121. doi: 10.1509/jm.15.0413

Wedel, M., and Pieters, R. (2014). "Looking at vision: eye/face/head tracking of consumers for improved marketing decisions," in *The Routledge Companion to the Future of Marketing*, eds L. Moutinho, E. Bigné, and A. K. Manrai (New York, NY: Routledge), 177–189.

Werkle-Bergner, M., Müller, V., Li, S. C., and Lindenberger, U. (2006). Cortical EEG correlates of successful memory encoding: implications for lifespan comparisons. *Neurosci. Biobehav. Rev.* 30, 839-854. doi: 10.1016/j.neubiorev. 2006.06.009

West, P. M., Brockett, P. L., and Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Mark. Sci.* 16, 370-391. doi: 10.1287/mksc.16.4.370

Zhou, R., Khemmarat, S., Gao, L., Wan, J., and Zhang, J. (2016). How YouTube videos are discovered and its impact on video views. *Multimed. Tools Appl.* 75, 6035-6058. doi: 10.1007/s11042-015-3206-0

Constructing a Reward-Related Quality of Life Statistic in Daily Life—a Proof of Concept Study Using Positive Affect

Simone J. W. Verhagen [1]*, Claudia J. P. Simons [1,2], Catherine van Zelst [1] and Philippe A. E. G. Delespaul [1,3]

[1] Department of Psychiatry and Neuropsychology, Faculty of Health Medicine and Lifesciences, Maastricht University, Maastricht, Netherlands, [2] GGzE Institute of Mental Health Care Eindhoven and De Kempen, Eindhoven, Netherlands, [3] Department of Adult Psychiatry, Mondriaan Mental Health Trust, Heerlen, Netherlands

Background: Mental healthcare needs person-tailored interventions. Experience Sampling Method (ESM) can provide daily life monitoring of personal experiences. This study aims to operationalize and test a measure of momentary reward-related Quality of Life (rQoL). Intuitively, quality of life improves by spending more time on rewarding experiences. ESM clinical interventions can use this information to coach patients to find a realistic, optimal balance of positive experiences (maximize reward) in daily life. rQoL combines the frequency of engaging in a relevant context (a 'behavior setting') with concurrent (positive) affect. High rQoL occurs when the most frequent behavior settings are combined with positive affect or infrequent behavior settings co-occur with low positive affect.

Methods: Resampling procedures (Monte Carlo experiments) were applied to assess the reliability of rQoL using various behavior setting definitions under different sampling circumstances, for real or virtual subjects with low-, average- and high contextual variability. Furthermore, resampling was used to assess whether rQoL is a distinct concept from positive affect. Virtual ESM beep datasets were extracted from 1,058 valid ESM observations for virtual and real subjects.

Results: Behavior settings defined by Who-What contextual information were most informative. Simulations of at least 100 ESM observations are needed for reliable assessment. Virtual ESM beep datasets of a real subject can be defined by Who-What-Where behavior setting combinations. Large sample sizes are necessary for reliable rQoL assessments, except for subjects with low contextual variability. rQoL is distinct from positive affect.

Conclusion: rQoL is a feasible concept. Monte Carlo experiments should be used to assess the reliable implementation of an ESM statistic. Future research in ESM should asses the behavior of summary statistics under different sampling situations. This exploration is especially relevant in clinical implementation, where often only small datasets are available.

Keywords: quality of life, experience sampling, reward, monte carlo simulations, behavior setting

INTRODUCTION

Mental health care is becoming more patient-centered. Every person is unique and the current classification systems miss relevant nuances for customized therapeutic interventions (Evans et al., 2013; McGorry and van Os, 2013). van Os (2014) argued to innovate assessment by making it more person-tailored and actively involve patients in the process. Clinicians should focus less on group characteristics and more on the individual's daily adaptation strategies and need for care (van Os, 2014). Over the years, diagnostic procedures have become more time consuming. Results are often complex latent structures that are not transparent and do not facilitate a collaborative communication between clinician and patient (van Staden, 2003; van Os, 2014). The main purpose of mental health care is to improve functioning as well as quality of life, in an empowering way. Most psychological interventions require motivation and engagement. An alienating communication does not help to engage patients in treatment. The ultimate goal is to assist patients in becoming more resilient, improve autonomy, reduce the impact of mental illness and improve well-being.

Resilient individuals are able to reduce their vulnerability by reducing the impact of symptoms and complaints in daily life. The reference point for therapeutic success is the actual moment-to-moment experience and functioning. The Experience Sampling Method (ESM) is a structured diary technique specially developed to appraise subjects in their daily interactions (Delespaul, 1995). The technique makes it possible to study subtle dynamic changes in momentary affective states that are difficult to assess in cross-sectional questionnaires. ESM has high ecological validity (Myin-Germeys et al., 2009; Trull and Ebner-Priemer, 2009) and allows person-tailored data collection. Since it reflects the subject's own daily life, the data facilitates transparent communication and collaborative care. ESM has been used for numerous research purposes in the general population (Jacobs et al., 2013), somatic health care (Parati et al., 2009) and mental health care (Walz et al., 2014). The method proved valuable in monitoring treatment effects (Munsch et al., 2009) and can be used as a treatment intervention (Wichers et al., 2011).

Quality of Life (QoL) is an important outcome indicator. Mental illness often decreases QoL, resulting in lowered subjective well-being and functioning (Fayers and Machin, 2013; Williams et al., 2015). QoL can be defined as "an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns" (The WHOQOL Group, 1995, p. 1405). Researchers and clinicians essentially consider QoL as a subjective concept. A broad operationalization combines different domains, such as social contact, physical health and environmental resources (The WHOQOL Group, 1995; Saxena et al., 1997). Affect, cognition, behavior and physical functioning influence experienced QoL (Spilker, 1990). The World Health Organization Quality of Life Assessment (WHOQOL) was developed to assess this comprehensive, multi-domain view of QoL (The WHOQOL Group, 1995). Other cross-sectional QoL measures either assess the impact of mental illness (Auquier et al., 2003; Fayers and Machin, 2013) or

monitor successes in treatment (Ruggeri et al., 2005; Yamauchi et al., 2008; Fleury et al., 2013). Structured interviews are used in the assessment of patients with a severe mental illness, both for subjective information (e.g., self-assessment) and objective information (e.g., social functioning) (Oliver et al., 1997; Priebe et al., 1999). Lehman (1983) showed that patients with severe mental illness were able to give an account of experienced QoL, similar to subjects from the general population.

Barge-Schaapveld et al. (1999) used ESM to study subjective well-being in daily life for patients with depression. They assumed that QoL varies from moment-to-moment and assessed momentary QoL (mQoL) repeatedly, using the question "In general, how is it going with you right now?" ESM questionnaires were administered 10 times a day for six consecutive days. Results confirmed between-subject and within-subject (temporal) variation in mQoL. Compared to healthy control subjects, depressed subjects reported lower mQoL, less activity, and experienced more negative affect and less positive affect. In addition, the variation of mQoL was higher in the depressed group. Furthermore, situational factors had a large influence on mQoL in both groups (Barge-Schaapveld et al., 1999). ESM is a feasible method for the assessment of momentary health-related QoL (Maes et al., 2015). QoL can be measured in the moment, under different daily life situations and varies between and within persons (Barge-Schaapveld et al., 1999; Maes et al., 2015). ESM is a compelling and comprehensive method because it assesses different factors that influence QoL, namely momentary affect and contextual variability.

To date, mQoL was an outcome statistic. Repeated self-assessments are made for a representative sample of moments during a specified period. Individual mQoL assessments are aggregated, yielding a statistic that represents pre or post intervention mQoL for a subject. The clinical relevance is limited, because these aggregated mQoL scores miss the necessary information to inform patients and clinicians on how to improve QoL dynamically over time. This requires an mQoL statistic that directly links treatment aims such as improving well-being with adaptation strategies in the moment and informs individuals and clinicians about choices in daily life.

The mechanism of reward or the process of reward seeking can be relevant to improve mQoL. Subjective well-being is related to reward experiences. Moreover, subjective well-being and reward-related neural activity are related (Gilleen et al., 2015). Reward experiences are the drivers of operant learning (Skinner, 1937). In operant conditioning, people learn from the consequences of their response and use that knowledge to guide future behavior choices (Skinner, 1937; Staddon and Cerutti, 2003). Stimulus-response associations are computed internally and updated frequently, allowing people to predict the outcome and choose responses from their repertoire accordingly. This implicit associative learning is driven by reinforcement. According to behavior theory, reinforcement strengthens or weakens the selection of behavior in a similar situation. Positive reinforcement occurs when a certain response to a new stimulus results in a valuated outcome and is thus rewarded, thereby increasing the likelihood of similar behavior in the future (Flora, 2004).

In mental health research, mechanisms of operant conditioning and reward are repeatedly linked to well-being. Lewinsohn (1974) hypothesized that depression is a consequence of low levels of response-contingent positive reinforcement. In a sample of college students, they correlated depressed mood with time spent on pleasant activities. Increased time in pleasant activities was viewed as an indicator of positive reinforcement. The results showed a moderately negative association: where pleasant activities decreased, experienced depression increased (Lewinsohn and Graf, 1973; Lewinsohn, 1974). A causal link between the two variables could not be established (Sweeney et al., 1982). Other studies have demonstrated positive affective experiences and reward experiences in relation to resilience in depressive subjects (Wichers et al., 2007, 2009). A randomized control trial (RCT) conducted by Geschwind et al. (2011) showed that mindfulness-based cognitive therapy aimed at increasing positive affect and the enjoyment of reward experiences during daily life was associated with a reduction in experienced depression.

Kramer et al. (2014) used an ESM-based intervention (ESM-I) to examine whether self-monitoring of positive affect (PA) is beneficial for depressive patients in addition to treatment as usual. ESM-derived feedback was used as a therapeutic tool to gain insight in implicit dynamic patterns that arise over time. Through person-tailored feedback sessions, hidden patterns were made explicit using visualization in graphs and figures (Kramer et al., 2014). Weekly ESM-I feedback sessions influenced the treatment of depression positively, with an effect still present at 6-month follow-up (Kramer et al., 2014). Contrary to these long-term effects, they found no significant impact of the ESM-I on daily experienced PA during the intervention or shortly after (Hartmann et al., 2015). Another RCT in young adults with depression also provides evidence that ESM-I may have positive impact on pleasure and PA by providing personalized lifestyle advice (van Roekel et al., 2017). Wichers et al. (2015) used ESM as a tool to prospectively observe implicit learning processes for reward-seeking and punishment-avoidant behavior in the context of daily life. They hypothesized that current behavior could be predicted by the experience of related behavior at previous time points. Results confirmed that affect moderates this association over time, both at beep and day level (Wichers et al., 2015).

Experience Sampling Method (ESM) is rooted in ecological psychology where contextual embeddedness is widely recognized (Heft, 2013). Barker introduced the term "behavior settings," to reflect the mutual relation between human behavior and the environment (Barker, 1965). Behavior settings are eco-behavioral entities which exist independent of persons and form self-regulating systems (Barker, 1968). They represent stable and identifiable constructs with both spatial and temporal indices and can provide opportunities or constrain the actions of persons (Barker, 1965; Wicker, 2012). The term behavior setting puts emphasis on processes and structures that often go unnoticed in the daily life of individuals (Wicker, 2012).

Imagine yourself on a market. Being there with friends will likely provide a different experience than being alone.

In addition, activities influence the experience. You could be working, buying a last ingredient for dinner or simply enjoying yourself. Other factors play a role as well, such as the location of the market, its attributes, the weather and time of day. All ingredients form the behavior setting—the rich and meaningful context.

Future research should emphasize the linkage between a person's affective experience and behavior setting characteristics, including the beliefs and know-how of this behavior setting (Wicker, 2012). With ESM, information on momentary affect can be gathered in the context of daily life (Delespaul, 1995). Behavior settings are important because they provide insight in the contextual variability of positive or negative affect. Clinicians can coach patients to engage in contexts (i.e., future behavior selection) to maximize individual patterns of positive affect and avoid negative affect, leading to more experienced QoL.

In line with the literature above, subjects can improve QoL by engaging more often and for longer time in affectively rewarding situations. Some may argue that extremely rewarding, low frequency behaviors can boost QoL (e.g., a holiday travel compensates for a boring job). However, exceptional situations do not rule moment-to-moment experiences, while frequent minor events that occur naturally in the flow of daily life have an impact on mental well-being (Peeters et al., 2003). The pursuit of reward cannot be a unidimensional focus. QoL does not increase linearly (more is better), but optimizes by balancing challenges of daily life. Even the most enjoyable job will be perceived differently when there is no time to relax anymore. Maximizing reward experiences in daily life, means spending most of our time in rewarding situations (eliciting high PA) and avoiding situations with low PA. Depressive patients, for instance, are out of balance and spend insufficient time on pleasant activities (Lewinsohn, 1974; Peeters et al., 2003; Thompson et al., 2012; Roekel et al., 2016). Their time budgets are not optimized and too much time is spent in behavior settings with low PA.

The reward-related QoL function (rQoL) reflects the momentary fit between context and optimal affective experience. ESM can monitor this process in treatment (Kramer et al., 2014). rQoL can be used in shared decision-making. Feeling good in some situations and bad in others is a transparent communication and most patients understand this intuitively. The collaborative ESM feedback sessions between patients and clinicians are the setting to discuss reward optimization. In a shared decision process, patients and clinicians select rewarding situations and explore how to increase the occurrence of well-being in daily life.

Can these situations be detected and personal profiles computed? When applied in clinical practice, does this improve the subject's overall well-being? This paper describes the development of a reward-related QoL function (rQoL) and the proof of concept of its applicability. A reward statistic is defined. It reflects individual daily life moment-to-moment variation in reward efficiency, by combining the actual (positive) affect with the occurrence frequency of the actual behavior setting. Data is collected using ESM. To our knowledge, this is the first study that uses mechanisms of reward to design a momentary quality of life measure.

METHOD

Sample

To assess the feasibility of the rQoL concept, ESM data from an existing dataset–the D-STIGMI study (van Zelst, 2014 p. 103) was used. The D-STIGMI study evaluates a psycho-educational coping skill training in people with severe mental illness using a RCT. The aim of the training was to increase resilience against stigmatization. The ESM data collection was an optional add-on to explore innovative outcome parameters. The Medical Ethics Committee of Maastricht University Medical Centre approved the study protocol under the number of NL3179406810. In the current study, only baseline ESM assessment data were used as seeds for the random simulated sets. ESM data were available for 27 participants.

Measurements

Experience Sampling Method

Experience Sampling Method (ESM) is a structured diary technique to assess moment-to-moment mental state changes in relation to situations in the daily life of individuals (Delespaul, 1995; Jacobs et al., 2011). The data of the D-STIGMI study was collected with the PsyMate™ device, a palm-top assessment tool developed for ESM data collection (http://www.psymate. eu/)[1]. The PsyMate™ was programmed to emit 10 beeps each day for six consecutive days. Beeps were generated at semi-random moments, within 90-min blocks, between 7.30 and 22.30 h. An auditory and visual signal indicated the availability of a short questionnaire. Responding lasted less than a minute. The questionnaires remained available for 15 min and subjects were instructed to promptly reply. Daily life experiences are captured in items assessing current affect, activities and context. Most items were rated on a 7-point Likert scale (ranging from 1 = "not at all" to 7 = "very"). A bipolar scale (−3 = negative, 0 = neutral, to 3 = positive) was used to assess stressful events. Context (activity, location and person present) and the use of substances were assessed with multi-optional checklists. The item "Overall, I feel well at the moment" was added to assess mQoL. Beeps are considered valid when the whole questionnaire is completed. Each subject could respond to a maximum of 60 beeps. In line with ESM guidelines, subjects were included in the analyses when they completed at least a third of the beeps (20 valid beeps) (Delespaul, 1995).

Assessment of rQoL

Momentary reward-related quality of life (rQoL) is defined as the fit between frequency of situations (behavior setting) and actual mental state. rQoL can be computed for each moment and provides feedback to clinicians and patients to collaboratively select intervention strategies that optimize the mental state in daily life and lead to more overall well-being. The related rQoL statistic uses the subject's own data to assess if he/she optimizes the selection of contexts to maximize positive mental states, meaning that rewarding situations should occur often and less rewarding situations should be avoided.

---

[1]Internet site: www.psymate.eu

The Ingredients of the Function

Mental states

The ESM questionnaire contains mood items that assess PA: I feel "cheerful," "satisfied," "relaxed," and "enthusiastic." Momentary PA was normalized by subject [$zPA_{(ij)} = PA_{(ij)} − PA_{(i)}$-with "i" for subjects and "j" for moments]. $zPA_{(ij)}$ yields positive scores for better than average mental states and negative scores for below average mental states. Better than average mental states are assumed to be rewarding.

Behavior setting

A meaningful situation is a behavior setting. Time, place, persons and activities define them. The ESM definition of a behavior setting uses the context information available in the beep-level questionnaires: the time of the day (morning, afternoon, evening), persons present ("with whom am I": "partner," "resident family," "family living away from home," "friends," "colleagues," "acquaintances," "strangers/others," and "nobody"), activity ("what am I doing": "resting," "work/school," "household/groceries," "hygiene," "eating/drinking," "relaxation," "doing something else" and "nothing") and location ("where am I": "home," "someone else's home," "work/school," "public place," "on the go," and "somewhere else"). This results in 3 × 8 × 8 × 6 = 1,152 potential combinations, of which many infrequently occur or never occur and result in empty cells for individual subjects. The detailed number of situations does not allow the selection of high and low frequency behavior settings for each subject. Therefore, the time of the day was omitted and the number of options for who, what and where limited to six each. For persons present, "partner" was included into "resident family" and "colleagues" into the "acquaintances" category. For activity, we combined "resting" with "doing nothing" and "household/groceries" with "hygiene." The options for location remained unchanged. The occurrence (as a proportion) of the 216 resulting behavior settings was computed for each individual subject. The cumulative proportion was computed with a break at 0.5 to differentiate the large set of infrequent situations and the much smaller set of frequent situations.

Reward function

The momentary rQoL statistic combines the frequency of the momentary behavior setting with the actual mental state. Specifically, the function uses the normalized positive affect score by individual [$zPA_{(ij)}$] and weights it with the frequency of occurrence of the individuals' behavior setting [$cp(BS_{ij})_i$]. Reward efficiency occurs when high frequency situations yield positive moods. High rQoL occurs when high frequency situations are combined with positive mental states or low frequency is limited to negative mental states. Low rQoL situations combine poor mental states with high frequency situations or elevated mental states with infrequent behavior settings. These characteristics are reflected in the formula:

$$rQoL\,(PA)_{ij} = zPA_{ij} \times \left(cp\left(BS_{ij} \right)_i − 0.5 \right)$$

in which:

rQoL(PA)$_{ij}$ is reward-based QoL computed on PA for subject i on moment j; zPA$_{ij}$ is the standardized PA score for subject i at

moment j; and $cp(BS_{ij})i$ is the cumulative proportion for subject i of occurrence of the current behavior setting for that subject i on moment j.

-0.5 is used to generate a cut-off score for high and low proportions. Other cut-off scores were explored (and are further explained in the analyses section).

Using this formula, a specific rQoL score was generated for each ESM moment. Negative scores (e.g., -0.1) represent low rQoL, whereas positive scores (e.g., 0.9) represent high rQoL. An alternative reward function can be defined with negative affect (NA) but PA was selected in line with other scholars in the field (Wichers et al., 2009, 2015). The rQoL statistic was computed using a small program written in a Stata^{TM} (v13.0) script.

Analyses

To assess the feasibility of the rQoL function, we assessed the impact of different choices using resampling methods (Monte Carlo experiments) to generate ESM data for virtual subjects. The sample size (N) is the number of ESM observations or beeps drawn from D-STIGMI dataset used as seeds. Different selections of contextual domains and different cut-off scores were explored to generate rQoL. Simulations were run using virtual subjects with data extracted from all beeps from the D-STIGMI seeding database. Because real subjects have more specific frequency distributions in behavior settings, virtual datasets were extracted for subjects with low, average and high situational diversity.

Which Sample Size Do We Need to Reliably Assess Behavior Setting in Individuals?

A Monte Carlo experiment was executed to study the effect of the number of available beeps on the average number of unique context combinations (behavior settings). Three different definitions for behavior setting were explored: a What only definition (BS_W: 6 situations), a What and Who–based definition (BS_WW: $6 \times 6 = 36$ situations) and a What, Who and Where definition (BS_WWW: $6 \times 6 \times 6 = 216$ combinations). Resampling made it possible to explore the alternative behavior setting definitions for different sampling sizes (20, 40, 60, 80, 100, 250, 500, 750, 1,000, 2,000, 3,000, 4,000, 8,000, 10,000 "valid" observations by virtual subject).

Which Sample Size Do We Need for Optimal Variation in Reward (rQoL)?

A second Monte Carlo experiment was executed to study the effect of available observations on the average scores and variation of rQoL, independently for BS_W, BS_WW and BS_WWW. This was done to provide insight in the number of observations needed to reliably generate rQoL. Initially, different cut-off scores (0.50, 0.40, 0.30, and 0.20) were explored to differentiate low and high frequency situations. Only the cut-off score 0.50 proved sensitive enough for meaningful rQoL detection and was therefore used in further analyses. Sample sizes that were explored, are 20, 40, 60, 80, 100, 250, 500, 750, 1,000, 2,000, 3,000, 4,000, 8,000, and 10,000 observations for the virtual subjects in the simulation.

What Is the Effect of the Actual Contextual Variation in Real Subjects?

In contrast to virtual subjects who live in the contexts of the group of individuals, actual subjects live in environments that are more restricted. We selected subjects with different levels of behavior setting differentiation (the number of non-empty BS_WWW categories) and ranked these individuals to compute percentiles. A Monte Carlo experiment was executed to simulate the rQoL functions for subjects with low (5th percentile), average (50th percentile) and high (95th percentile) variability in contextual domains. Sample sizes that were explored are 20, 40, 60, 80, and 100 "valid" observations by virtual subject.

Is Reward-Related rQoL Something Different than Positive Affect?

A final Monte Carlo experiment was executed to assess whether momentary PA and rQoL were separate concepts. Pearson's product-moment correlations between PA scores and rQoL scores were computed for each resampled set of momentary data, using varying sampling sizes. This was done separately for BS_WWW, BS_WW and BS_W in the overall sample.

The seeding data for the Monte Carlo experiments were selected using ESM observations from real subjects combined together as the sampling domain. When not enough unique empirical data were available (simulated samples exceeding available observations), we sampled with replacement. For each simulation, 1000 samples were drawn. Analyses were performed in Stata^{TM} (v13.0). The do-file is added in the Supplementary Material.

RESULTS

Subject Characteristics

Twenty-seven patients with a severe mental illness were included in the ESM baseline measurement of the D-STIGMI study. Four patients had insufficient valid beeps ($<1/3$), thus 23 patients were included as seeds for Monte Carlo experiments. The sampling set includes 1058 valid beeps, at average 48 per subject ($SD = 9.03$, range 22–63). No significant differences were found between the original sample and the seed-sample on age ($p = 0.97$) and sex ($p = 0.53$). Demographic information is summarized in **Table 1**.

Which Sample Size Do We Need to Assess Behavior Settings in Individuals Reliably?

Figure 1 shows the number of unique behavior settings generated by the resampling procedures for different sample sizes. The theoretical number of behavior settings was 216 but the beeps actually only contained 113 options (the empirical ceiling). The graph has three phases: from $N = 20$ (mean = 13; SD = 1.75) to $N = 100$ (mean = 33; SD = 3.31), from $N = 100$ to $N = 1,000$ (mean = 92; SD = 3.6) and from $N = 1,000$ to $N = 10,000$ (mean = 113; SD = 0.16). Around $N = 8,000$ (mean = 113; SD = 0.16), saturation is reached. Standard errors are low in small samples (SEM = 0.39; $N = 20$) and increase over the second phase (max SEM = 4.08; $N = 500$) and finally to reduce again (SEM = 0.002; $N = 10,000$). **Figures 1B,C** reflect simulations for the BS_WW (theoretical $6 \times 6 = 36$ options, sample 36 options) and BS_W

TABLE 1 | Demographics and characteristics of the 23 participants seed-sample.

Characteristics	N (%)	Mean	SD	Range	Details
Age (years)		46	9.95	29-62	
GENDER					
Female	60				
Male	40				
DSM-IV DIAGNOSTICS					
Single diagnosis	61				
Multiple diagnoses	35				
Missing diagnostics	4				
CLINICAL CLASSIFICATION					
Primary psychotic vulnerability	68				(292.12/295.30/295.70/298.9/296.40)
Primary non-psychotic vulnerability	32				(296.31/299.80/301.83/301.83/309.81)
GAF symptoms		46.45	14.27	28–70	
GAF handicap		45.00	9.83	31–70	
BEEP CHARACTERISTICS (N = 1,058)					
Positive affect (overall beeps)		4.00	1.41	1–7	
Positive affect (by subject)		4.00	0.99	1.3–6.3	
NUMBER OF CONTEXT COMBINATIONS					
Unique BS_WWW combinations	113				Possible = 216
Unique BS_WW combinations	36				Possible = 36
Unique BS_W combinations	6				Possible = 6
CONTEXT COMBINATIONS PER SUBJECT					
Low BS variability subject	5				
Average BS variability subject	17				
High BS variability subject	28				

292.12 = Induced psychotic disorder, with hallucinations. 295.3 = Schizophrenia: Paranoid type. 295.7 = Schizoaffective disorder. 298.9 = Psychotic disorder NOS. 296.4 = Bipolar I disorder, most recent episode hypomanic. 296.31 = Major depressive disorder, mild. 299.8 = Rett's disorder; Asperger's disorder; Pervasive developmental disorder NOS. 301.83 = Borderline personality disorder. 309.81 = posttraumatic stress disorder.
GAF, Global Assessment of Functioning; BS, Behavior Setting.

behavior settings (theoretical 6 options, sample 6 options). The same pattern replicates but the ceiling is reached with smaller samples (around 500 observations in the BS_WW and 100 in the BS_W situation).

Which Sample Size Do We Need for Optimal Variation in Reward Quality of Life (rQoL)?

For different sampling sizes, the rQoL was computed at each moment of the simulated virtual subject. As expected, all samples had an average of 0.00, the neutral point of the rQoL function. Standard errors of the mean were low and are considered negligible ($SEM = 0.06$ to $SEM = 0.003$). For BS_WWW and BS_WW (**Figures 2A,B**), the range of rQoL scores increased up to sample sizes of 500 beeps. For BS_W the range was more restricted but reached its maximum for sampling sizes of 60.

What Is the Effect of the Actual Contextual Variation in Real Subjects?

The selected subjects (low-, average-, high variability in behavior setting) responded reliably to respectively 31, 45, and 55 beeps, with respectively 5, 17, and 28 different behavior settings (using the BS_WWW combination). Results are presented in **Figure 3**.

Part 3a shows the increase in average rQoL scores for the subject with low behavior variability (p5; 5th percentile), average behavior setting variability (p50; 50th percentile) and high behavior setting variability (p95; 95th percentile). Increases in sampling size did not affect the scores for subjects with low behavior variation (1.1 mean difference), but did for subjects with high variation (12.6 mean difference). Part 3b shows the range of minimum and maximum scores. This confirms the previous observation. Smaller sampling sizes are possible in subjects living restricted lives.

Is Reward-Related QoL Something Different than Positive Affect?

The Pearson's product-moment correlations of the Monte Carlo experiments on sampled sets of PA scores and rQoL are summarized in **Figure 4**. Looking at the average correlation scores, weak associations were found between PA and rQoL for BS_WWW, BS_WW, and BS_W [range $r_{(18)} = 0.11$, $p < 0.001$ to $r_{(9, 998)} = 0.005$, $p = < 0.001$]. The range between minimum and maximum correlational scores in all three variations of behavior setting is large in smaller sample sizes (highest difference from $min = -0.91$ to $max = 0.90$) and decreases with larger sample sizes (lowest difference from $min = -0.02$ to $max = 0.04$).
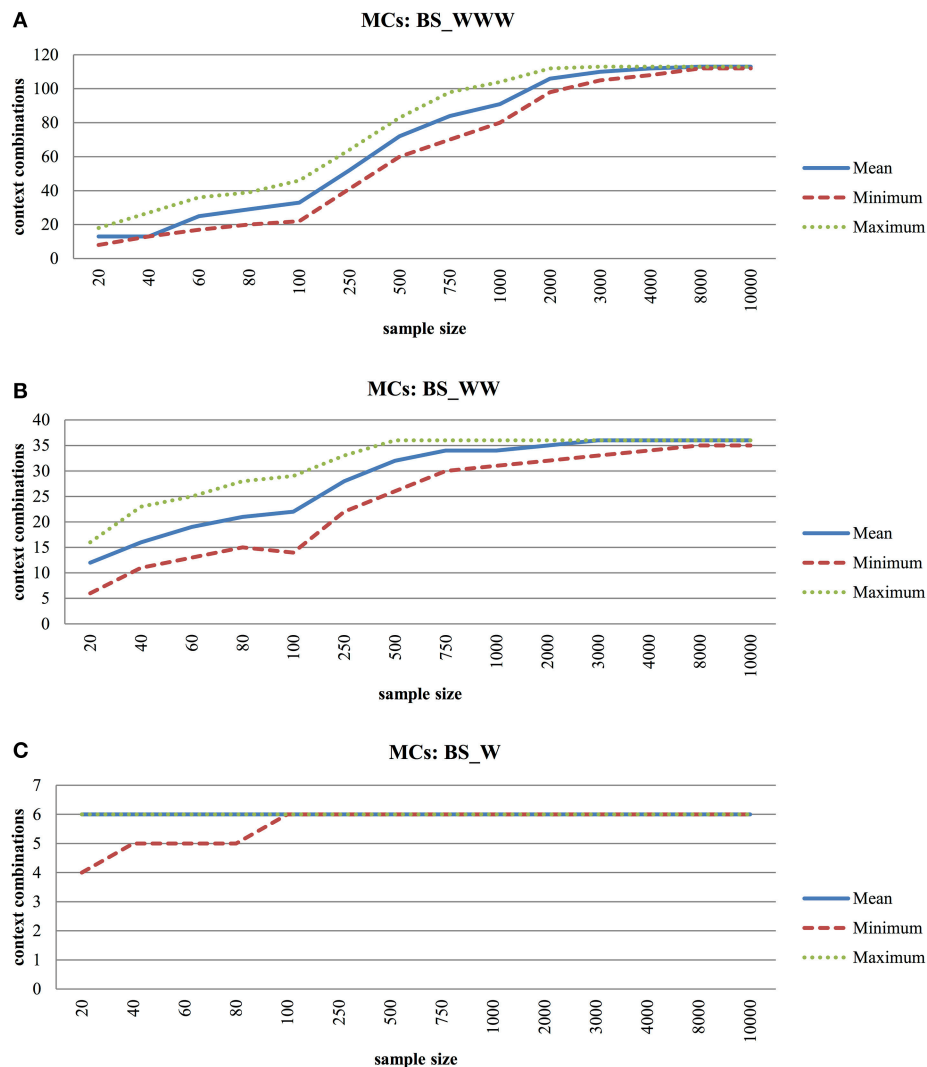
FIGURE 1 | (A) Monte Carlo Experiment (MCs) to Explore Average-, Minimum-, and Maximum Number of Unique Context Combinations with Increased Sample Size for Behavior Settings Including Who, What, Where Information (BS_WWW). **(B)** Monte Carlo Experiment (MCs) to Explore Average-, Minimum- and Maximum Number of Unique Context Combinations with Increased Sample Size for Behavior Settings Including Who, What Information (BS_WW). **(C)** Monte Carlo Experiment (MCs) to Explore Average-, Minimum- and Maximum Number of Unique Context Combinations with Increased Sample Size for Behavior Settings Including What Information (BS_W).

DISCUSSION

General Conclusion

The purpose of the current proof-of-concept study is two-fold. First, a momentary rQoL statistic was defined. Second, Monte Carlo experiments were performed with samples of beep-level data for virtual subjects to check the feasibility and initial validity of the statistic. Momentary rQoL integrates affective experience (positive affect) and situations (behavior setting). A cut-off score of 0.50 of the cumulative proportion was chosen to separate low and high frequency situations. This proved to be the best choice to detect relevant contexts for reward efficiency in most subjects. Results show that the rQoL statistic is feasible. The rQoL statistic is defined at the moment level, allowing assessment of

change over time. At a specific moment in time, a positive rQoL score indicates good reward efficiency, whereas a negative score indicates bad reward efficiency.

Persons present (Who), activity (What) and location (Where) were used to define different conceptualizations of a behavior setting, namely Who-What-Where (BS_WWW), Who-What (BS_WW), and What (BS_W). For Who-What-Where combinations, not all theoretical possibilities were available in the reference beep dataset. This dataset included 113 unique combinations out of the 216 theoretical options. Some situations had low frequencies (e.g. working with your partner or doing household/groceries at work) or simply did not occur in the group of patients with a severe mental illness. First, Monte Carlo experiments were performed to check which
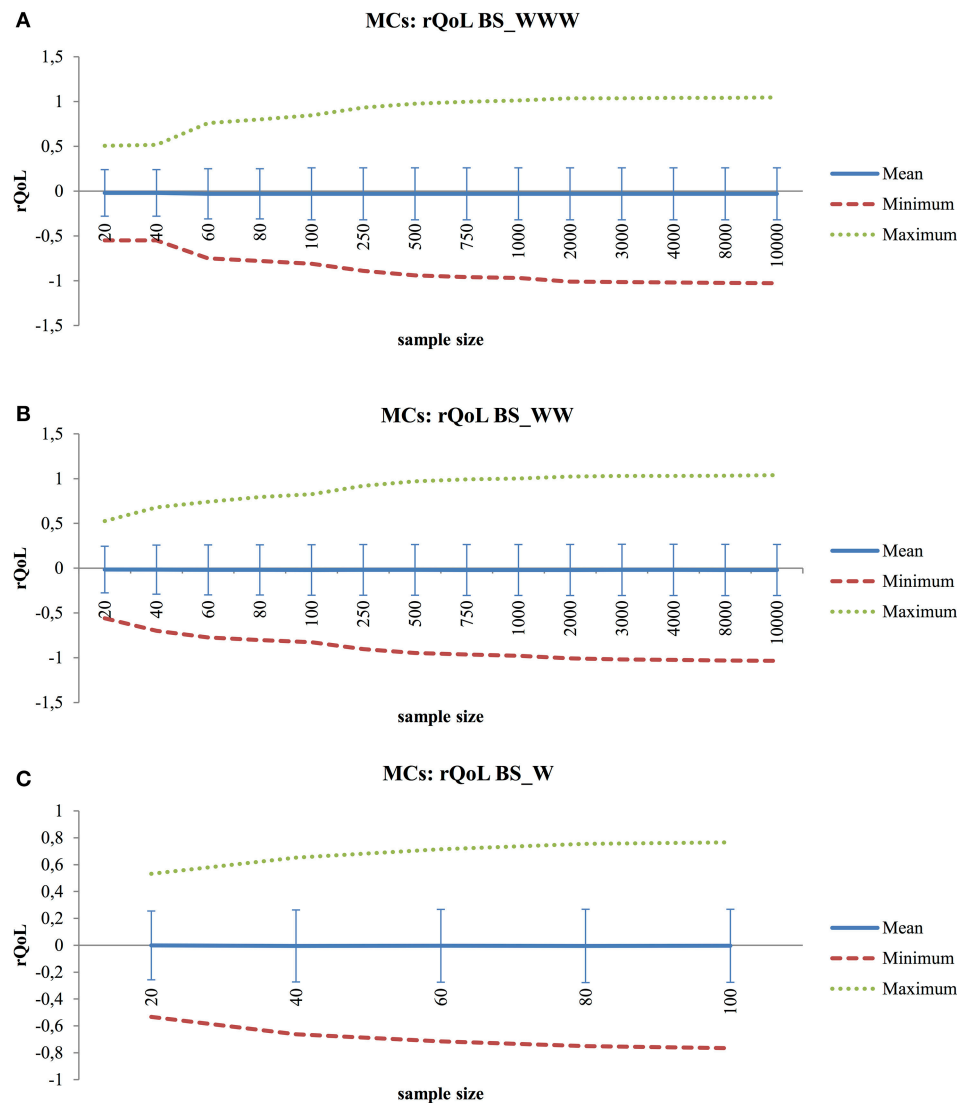
FIGURE 2 | (A) Monte Carlo Experiment (MCs) to Explore Average-, Minimum-, Maximum- and Standard Deviation Scores of rQoL(PA) with Increased Sample Size for Behavior Settings Including Who, What, Where Information (BS_WWW). **(B)** Monte Carlo Experiment (MCs) to Explore Average-, Minimum-, Maximum-, and Standard Deviation Scores of rQoL(PA) with Increased Sample Size for Behavior Settings Including Who, What Information (BS_WW). **(C)** Monte Carlo Experiment (MCs) to Explore Average-, Minimum-, Maximum-, and Standard Deviation Scores of rQoL(PA) with Increased Sample Size for Behavior Settings Including What Information (BS_W).

sample size is needed for reliable behavior setting calculations in individuals. Overall, large sample sizes were needed to generate realistic frequency distributions (time budgets) in virtual subjects (saturation was reached at $N = 8,000$ for BS_WWW, $N = 500$ for BS_WW, and $N = 100$ for BS_W). Next, Monte Carlo experiments were performed to explore which sample size is needed to detect optimal rQoL variation. The range of rQoL scores increased up to generated samples of 500 beeps for BS_WWW and BS_WW, whereas the limit was reached at 60 beeps for BS_W. Therefore, a behavior setting defined by Who-What proved most useful: it provides sufficient behavior setting variation to generate a reliable frequency distribution at acceptable sample sizes. An optimal spread in rQoL variation

is obtained after collecting 500 observations, meaning that all possible behavior-setting combinations are present in the sampling period (see **Figure 2B**). However, a minimum of 100 beeps is required to reliably calculate the rQoL statistic. The sample sizes needed for BS_WWW are unrealistic in ESM. These results suggest that an extended sampling period is needed before the rQoL statistic can be integrated as an active part of treatment.

Further, Monte Carlo experiments were used to explore actual situational variations in real subjects with low, average and high behavior setting variation with combinations of persons present, activity and location as behavior setting. This behavior setting definition was chosen because it theoretically provides the largest chance of finding a decent spread in unique situations
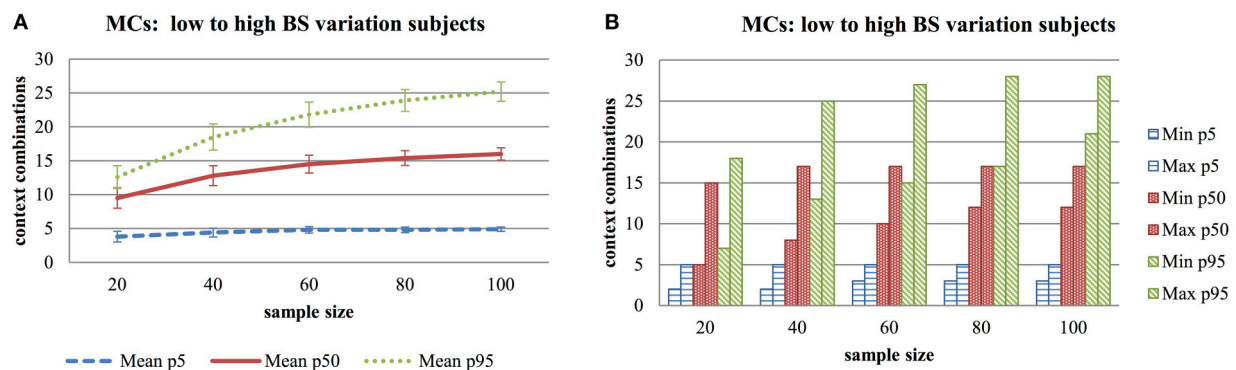
FIGURE 3 | (A) Monte Carlo Experiment (MCs) to Explore Average- and Standard Deviations Scores of Unique Context Combinations for Subjects with Low-, Medium- and High Diversity in Behavioral Setting, with Increased Sample Size for Behavior Settings Including Who, What, Where Information (BS_WWW). **(B)** Monte Carlo Experiment (MCs) to Explore Minimum- and Maximum Number of Unique Context Combinations for Subjects with Low-, Medium- and High Diversity in Behavioral Setting, with Increased Sample Size for Behavior Settings Including Who, What, Where Information (BS_WWW).

in this patient group. Results show that small sample sizes are possible in subjects with low behavior setting variations ($N > 40$), whereas larger sample sizes ($N > 100$) are needed for subjects with average and high behavior setting variations. It is difficult to assess which definition of behavior setting is sufficient for individual subjects. For now, behavior settings defined by Who-What-Where combinations seem the best option because the restricted living environments of individual subjects results in less overall situational variation.

Finally, Monte Carlo experiments were run to see whether momentary rQoL is distinct from PA. Only weak correlations were found and results confirmed that the concepts assess different aspects of daily life mental states. Additionally, the momentary rQoL scores were correlated with "In general, how is it going with you right now" (mQoL). For this, the overall beep sample of 1058 valid observations was used. Results show a moderate positive correlation [$r_{(1, 056)} = 0.33$, $p < 0.001$], with mQoL explaining 11 percent of the variation in rQoL.

Strengths

To our knowledge, this is the first study that combines affective experience with behavior settings in an operationalization of rQoL in daily life. A main strength is the use of ESM data collected in the flow of daily life, making the rQoL statistic highly specific to the situation of the subject. Monte Carlo experiments are especially suited for exploration of properties and sampling characteristics of specific combined statistics (Mooney, 1997). Researchers compute different parameters, but insufficiently realize the relation and biases due to sampling characteristics. Knowing how a statistic responds to different sampling characteristics is particularly important when applied in treatment.

LIMITATIONS

The choices made to operationalize momentary rQoL can seem arbitrary. Other options could be explored. For example, the

rQoL statistic can be computed with negative affect. Additional cut-off scores could be explored to differentiate low and high frequency situations. The behavior settings are defined on persons present, activity and location, although more aspects of the environment could be relevant. Behavior settings are complex entities which include a number of contextual variables (Barker, 1965). Here we excluded, for example, temporal indices such as the time of the day and limited the categorical options so that a critical mass of workable data remains. With advances in technology, other factors such as heart rate or weather reports could be more easily combined with ESM data, thereby increasing the accuracy of the behavior setting. In the future, it is possible to harvest big data, such as GPS location, sensor data, or geo-political events to enrich the situational information without increasing subject burden. However, the main purpose of this study was to explore a first operationalization based on available ESM data and test the behavior of the momentary rQoL statistic using Monte Carlo experiments.

Furthermore, the generated samples using real subject data (for the analysis of subjects with low-, average- and high behavior setting variability) were oversaturated and the same records were used repeatedly (due to replacement). The used sample included insufficient subjects with large beep datasets (>1000 observations). It would be interesting to replicate these analyses when longer series become available.

Another limitation is the use of a specific sample of patients with severe mental illness as seeds for the Monte Carlo experiments. These subjects often lead restricted lives with limited variation in daily life activities (Lewinsohn, 1974; Holloway and Carson, 2002). It would be interesting to see how the frequencies in behavior setting differ between these patients and the general population. This first operationalization was made from a clinical perspective, to explore the possibility of an rQoL statistic that is meaningful to patients with severe mental illness. However, the basic components of the rQoL statistic could be relevant across clinical populations. For example, patients with depression or anxiety could also use the statistic within
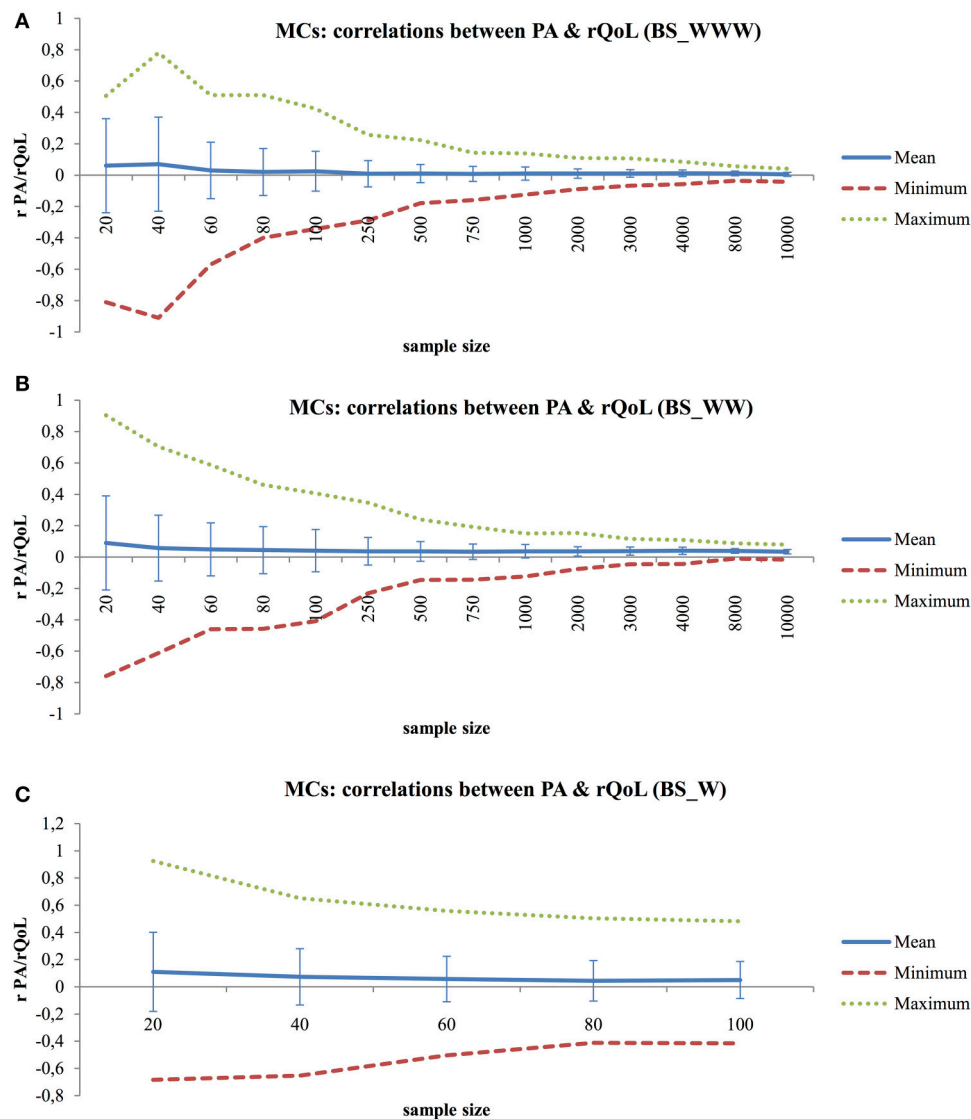
FIGURE 4 | (A) Monte Carlo Experiment (MCs) to Explore the Relationship between PA Scores and rQoL Scores with Increased Sample Size for Behavior Settings Including Who, What, Where Information (BS_WWW), Using Pearson's Product-Moment Correlations. **(B)** Monte Carlo Experiment (MCs) to Explore the Relationship between PA Scores and rQoL Scores with Increased Sample Size for Behavior Settings Including Who, What Information (BS_WW), Using Pearson's Product-Moment Correlations. **(C)** Monte Carlo Experiment (MCs) to Explore the Relationship between PA Scores and rQoL Scores with Increased Sample Size for Behavior Settings Including What Information (BS_W), Using Pearson's Product-Moment Correlations.

treatment to optimize their balance in rQoL. Future research should use Monte Carlo simulations on ESM data collected in other populations; to see how the frequency in behavior setting is distributed, to calculate which sample size would work best, and to see in what situation the rQoL statistic can be meaningfully calculated. It is conceivable that situational variability differs between populations (and between individuals).

Implications and Further Research

This proof-of-concept study indicates that momentary rQoL is a feasible statistic. Monte Carlo experiments provide valuable insight in the behavior of the statistic under different sampling restraints. The methodology can be used to further

improve rQoL. Monte Carlo experiments should be used more frequently in ESM studies. Several suggestions were made for future research. The question remains whether reward-related optimized well-being is actually quality of life; maybe a better description is more adequate. The link between the rQoL statistic and mQoL could further be explored, as well as the relation of rQoL with other (cross-sectional) measures of QoL. Furthermore, the statistic should be explored in other populations that engage in more diversified behavior settings.

It is interesting to explore whether the rQoL statistic improves targeted communication with patients. The statistic could be used to identify situations that result in low rQoL, so that changes can be made in daily life and progress can be monitored (see

the Supplementary Material for a hypothetical case example). A previous ESM-based feedback intervention (Kramer et al., 2014) improved therapeutic outcome. The question remains if well-being can be improved by a person-tailored rQoL feedback intervention that monitors reward experiences in daily life. Shared decision-making is facilitated when clinicians and patients share the same information. ESM data, disclosed by smart feedback, can provide this context. By integrating the clinician's expertise with the goals and knowledge of patients and relatives, and by looking at environmental daily life challenges and opportunities, more suggestions that are realistic can be made for optimizing reward in daily life, possibly leading to improved well-being and QoL.

AUTHOR CONTRIBUTIONS

SV and PD worked on the conceptualization of rQoL and subsequent analyses. SV wrote the initial article. All authors (SV, PD, CS, and CvZ) provided substantial feedback to the rationale of the article and contributed to later versions of the article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01917/full#supplementary-material

REFERENCES

Auquier, P., Simeoni, M. C., Sapin, C., Reine, G., Aghababian, V., Cramer, J., et al. (2003). Development and validation of a patient-based health-related quality of life questionnaire in schizophrenia: the S-QoL. *Schizophr. Res.* 63, 137–149. doi: 10.1016/S0920-9964(02)00355-9

Barge-Schaapveld, D. Q., Nicolson, N. A., and Berkhof, J. (1999). Quality of life in depression: daily life determinants and variability. *Psychiatry Res.* 88, 173–189. doi: 10.1016/S0165-1781(99)00081-5

Barker, R. G. (1965). Explorations in ecological psychology. *Am. Psychol.* 20:1. doi: 10.1037/h0021697

Barker, R. G. (1968). *Ecological psychology: Concepts and Methods for Studying the Environment of Human Behavior.* Stanford, CA: Stanford University Press.

Delespaul, P. (1995). *Assessing Schizophrenia in Daily Life: The Experience Sampling Method.* Maastricht, Universitaire Pers Maastricht.

Evans, S. C., Reed, G. M., Roberts, M. C., Esparza, P., Watts, A. D., Correia, J. M., et al. (2013). Psychologists' perspectives on the diagnostic classification of mental disorders: results from the WHO-IUPsyS Global Survey. *Int. J. Psychol.* 48, 177–193. doi: 10.1080/00207594.2013.804189

Fayers, P. M., and Machin, D. (2013). *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes.* West Sussex: John Wiley & Sons.

Fleury, M. J., Grenier, G., Bamvita, J. M., Tremblay, J., Schmitz, N., and Caron, J. (2013). Predictors of quality of life in a longitudinal study of users with severe mental disorders. *Health Q. Life Outcomes* 11, 92–92. doi: 10.1186/1477-7525-11-92

Flora, S. R. (2004). *The Power of Reinforcement.* Albany, NY: State University of New York Press.

Geschwind, N., Peeters, F., Drukker, M., van Os, J., and Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: a randomized controlled trial. *J. Consult. Clin. Psychol.* 79, 618–628. doi: 10.1037/a0024595

Gilleen, J., Shergill, S. S., and Kapur, S. (2015). Impaired subjective well-being in schizophrenia is associated with reduced anterior cingulate activity during reward processing. *Psychol. Med.* 45, 589–600. doi: 10.1017/S0033291714001718

Hartmann, J. A., Wichers, M., Menne-Lothmann, C., Kramer, I., Viechtbauer, W., Peeters, F., et al. (2015). Experience sampling-based personalized feedback and positive affect: a randomized controlled trial in depressed patients. *PLoS ONE* 10:e0128095. doi: 10.1371/journal.pone.0128095

Heft, H. (2013). An ecological approach to psychology. *Rev. Gen. Psychol.* 17:162. doi: 10.1037/a0032928

Holloway, F., and Carson, J. (2002). Quality of life in severe mental illness. *Int. Rev. Psychiatry* 14, 175–184. doi: 10.1080/09540260220145000

Jacobs, N., Menne-Lothmann, C., Derom, C., Thiery, E., van Os, J., and Wichers, M. (2013). Deconstructing the familiality of variability in momentary negative and positive affect. *Acta Psychiatr. Scand.* 127, 318–327. doi: 10.1111/j.1600-0447.2012.01924.x

Jacobs, N., van OS, J., Derom, C., Thiery, E., Delespaul, P., and Wichers, M. (2011). Neuroticism explained? From a non-informative vulnerability marker to informative person–context interactions in the realm of daily life. *Br. J. Clin. Psychol.* 50, 19–32. doi: 10.1348/014466510X491397

Kramer, I., Simons, C. J., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., et al. (2014). A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry* 13, 68–77. doi: 10.1002/wps.20090

Lehman, A. F. (1983). The well-being of chronic mental patients: assessing their quality of life. *Arch. Gen. Psychiatry* 40, 369–373. doi: 10.1001/archpsyc.1983.01790040023003

Lewinsohn, P. M. (1974). "A behavioral approach to depression," in *The Psychology of Depression: Contemporary Theory and Research,* eds R. J. Friedman and M. M. Katz (Washington, DC: Winston-Wiley).

Lewinsohn, P. M., and Graf, M. (1973). Pleasant activities and depression. *J. Consult. Clin. Psychol.* 41:261. doi: 10.1037/h0035142

Maes, I. H., Delespaul, P. A., Peters, M. L., White, M. P., van Horn, Y., Schruers, K., et al. (2015). Measuring health-related quality of life by experiences: the experience sampling method. *Value Health* 18, 44–51. doi: 10.1016/j.jval.2014.10.003

McGorry, P., and van Os, J. (2013). Redeeming diagnosis in psychiatry: timing versus specificity. *Lancet* 381, 343–345. doi: 10.1016/S0140-6736(12)61268-9

Mooney, C. Z. (1997). *Monte Carlo Simulation.* Sage: Publications.

Munsch, S., Meyer, A. H., Milenkovic, N., Schlup, B., Margraf, J., and Wilhelm, F. H. (2009). Ecological momentary assessment to evaluate cognitive-behavioral treatment for binge eating disorder. *Int. J. Eat. Disord.* 42, 648–657. doi: 10.1002/eat.20657

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., and van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychol. Med.* 39, 1533–1547. doi: 10.1017/S0033291708004947

Oliver, J. P., Huxley, P. J., Priebe, S., and Kaiser, W. (1997). Measuring the quality of life of severely mentally ill people using the lancashire quality of life profile. *Soc. Psychiatry Psychiatr. Epidemiol.* 32, 76–83. doi: 10.1007/BF00788924

Parati, G., Omboni, S., and Bilo, G. (2009). Why is out-of-office blood pressure measurement needed? *Hypertension* 54, 181–187. doi: 10.1161/HYPERTENSIONAHA.108.122853

Peeters, F., Nicolson, N. A., Berkhof, J., Delespaul, P., and deVries, M. (2003). Effects of daily events on mood states in major depressive disorder. *J. Abnorm. Psychol.* 112:203. doi: 10.1037/0021-843X.112.2.203

Priebe, S., Huxley, P., Knight, S., and Evans, S. (1999). Application and results of the Manchester Short Assessment of Quality of Life (MANSA). *Int. J. Soc. Psychiatry* 45, 7–12. doi: 10.1177/002076409904500102

Roekel, E., Bennik, E. C., Bastiaansen, J. A., Verhagen, M., Ormel, J., Engels, R. C., et al. (2016). Depressive symptoms and the experience of pleasure in daily life: an exploration of associations in early and late adolescence. *J. Abnorm. Child Psychol.* 44, 999–1009. doi: 10.1007/s10802-015-0090-z

Ruggeri, M., Nosè, M., Bonetto, C., Cristofalo, D., Lasalvia, A., Salvi, G., et al. (2005). Changes and predictors of change in objective and subjective quality of life: multiwave follow-up study in community psychiatric practice. *Br. J. Psychiatry* 187, 121–130. doi: 10.1192/bjp.187.2.121

Saxena, S., Orley, J., and Group, W. (1997). Quality of life assessment: the World Health Organization perspective. *Eur. Psychiatry* 12, 263s–266s. doi: 10.1016/S0924-9338(97)89095-5

Skinner, B. F. (1937). Two types of conditioned reflex: a reply to Konorski and Miller. *J. Gen. Psychol.* 16, 272–279. doi: 10.1080/00221309.1937.9917951

Spilker, B. (1990). *Quality of Life Assessments in Clinical Trials.* New York, NY, Raven Press.

Staddon, J. E. R., and Cerutti, D. T. (2003). Operant conditioning. *Annu. Rev. Psychol.* 54, 115–144. doi: 10.1146/annurev.psych.54.101601.145124

Sweeney, P. D., Shaeffer, D. E., and Golin, S. (1982). Pleasant events, unpleasant events, and depression. *J. Pers. Soc. Psychol.* 43, 136–144. doi: 10.1037/0022-3514.43.1.136

The WHOQOL Group (1995). The World Health Organization quality of life assessment (WHOQOL): position paper from the World Health Organization. *Soc. Sci. Med.* 41, 1403–1409.

Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Gotlib, I. H. (2012). The everyday emotional experience of adults with major depressive disorder: examining emotional instability, inertia, and reactivity. *J. Abnorm. Psychol.* 121:819. doi: 10.1037/a0027978

Trull, T. J., and Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *Psychol. Assess.* 21, 457–462. doi: 10.1037/a0017653

van Os, J. (2014). *De Dsm-5 voorbij! Persoonlijke Diagnostiek in een nieuwe GGZ.*, Leusden: Diagnosis Uitgevers.

van Roekel, E., Vrijen, C., Heininga, V. E., Masselink, M., Bos, E. H., and Oldehinkel, A. J. (2017). An exploratory randomized controlled trial of personalized lifestyle advice and tandem skydives as a means to reduce anhedonia. *Behav. Ther.* 48, 76–96. doi: 10.1016/j.beth.2016.09.009

van Staden, W. C. (2003). Diagnostic understanding and diagnostic explanation in psychiatry. *Curr. Opin. Psychiatry* 16, 667–671. doi: 10.1097/00001504-200311000-00011

van Zelst, C. J. A. M. (2014). *Inside Out: On Stereotype Awareness, Childhood Trauma and Stigma in Psychosis.* Doctoral dissertation, Maastricht University.

Walz, L. C., Nauta, M. H., and Aan Het Rot, M. (2014). Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: a systematic review. *J. Anxiety Disord.* 28, 925–937. doi: 10.1016/j.janxdis.2014.09.022

Wichers, M. C., Barge-Schaapveld, D. Q., Nicolson, N. A., Peeters, F., de Vries, M., and Mengelers, R. et al. (2009). Reduced stress-sensitivity or increased reward experience: the psychological mechanism of response to antidepressant medication. *Neuropsychopharmacology* 34, 923–931. doi: 10.1038/npp.2008.66

Wichers, M. C., Myin-Germeys, I., Jacobs, N., Peeters, F., Kenis, G., Derom, C., et al. (2007). Evidence that moment-to-moment variation in positive emotions buffer genetic risk for depression: a momentary assessment twin study. *Acta Psychiatr. Scand.* 115, 451–457. doi: 10.1111/j.1600-0447.2006.00924.x

Wichers, M., Hartmann, J. A., Kramer, I. M., Lothmann, C., Peeters, F., van Bemmel, L., et al. (2011). Translating assessments of the film of daily life into person-tailored feedback interventions in depression. *Acta Psychiatr. Scand.* 123, 402–403. doi: 10.1111/j.1600-0447.2011.01684.x

Wichers, M., Kasanova, Z., Bakker, J., Thiery, E., Derom, C., Jacobs, N., et al. (2015). From affective experience to motivated action: tracking reward-seeking and punishment-avoidant behaviour in real-life. *PLoS ONE* 10:e0129722. doi: 10.1371/journal.pone.0129722

Wicker, A. W. (2012). Perspectives on behavior settings: with illustrations from Allison's Ethnography of a Japanese Hostess club. *Environ. Behav.* 44, 474–492. doi: 10.1177/0013916511398374

Williams, E., Sands, N., Elsom, S., and Prematunga, R. K. (2015). Mental health consumers' perceptions of quality of life and mental health care. *Nurs. Health Sci.* 17, 299–306. doi: 10.1111/nhs.12189

Yamauchi, K., Aki, H., Tomotake, M., Iga, J., Numata, S., Motoki, I., et al. (2008). Predictors of subjective and objective quality of life in outpatients with schizophrenia. *Psychiatry Clin. Neurosci.* 62, 404–411. doi: 10.1111/j.1440-1819.2008.01818.x

Network Approach to Understanding Emotion Dynamics in Relation to Childhood Trauma and Genetic Liability to Psychopathology: Replication of a Prospective Experience Sampling Analysis

Laila Hasmi[1]*, Marjan Drukker[1], Sinan Guloksuz[1,2], Claudia Menne-Lothmann[1], Jeroen Decoster[3], Ruud van Winkel[1,3], Dina Collip[1], Philippe Delespaul[1], Marc De Hert[3], Catherine Derom[4,5], Evert Thiery[6], Nele Jacobs[1,7], Bart P. F. Rutten[1], Marieke Wichers[8] and Jim van Os[1,9,10]

[1] Department of Psychiatry and Psychology, Maastricht University Medical Centre, Maastricht, Netherlands, [2] Department of Psychiatry, Yale School of Medicine, New Haven, CT, United States, [3] University Psychiatric Centre KU Leuven, Leuven, Belgium, [4] Centre of Human Genetics, University Hospitals Leuven, KU Leuven, Leuven, Belgium, [5] Department of Obstetrics and Gynecology, Ghent University Hospitals, Ghent University, Ghent, Belgium, [6] Department of Neurology, Ghent University Hospital, Ghent University, Ghent, Belgium, [7] Faculty of Psychology and Educational Sciences, Open University of the Netherlands, Heerlen, Netherlands, [8] Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion Regulation, University of Groningen, University Medical Center Groningen, Groningen, Netherlands, [9] Department of Psychosis Studies, Institute of Psychiatry, King's Health Partners, King's College London, London, United Kingdom, [10] Department of Psychiatry, Brain Centre Rudolf Magnus, University Medical Centre Utrecht, Utrecht, Netherlands

Background: The network analysis of intensive time series data collected using the Experience Sampling Method (ESM) may provide vital information in gaining insight into the link between emotion regulation and vulnerability to psychopathology. The aim of this study was to apply the network approach to investigate whether genetic liability (GL) to psychopathology and childhood trauma (CT) are associated with the network structure of the emotions "cheerful," "insecure," "relaxed," "anxious," "irritated," and "down"—collected using the ESM method.

Methods: Using data from a population-based sample of twin pairs and siblings (704 individuals), we examined whether momentary emotion network structures differed across strata of CT and GL. GL was determined empirically using the level of psychopathology in monozygotic and dizygotic co-twins. Network models were generated using multilevel time-lagged regression analysis and were compared across three strata (low, medium, and high) of CT and GL, respectively. Permutations were utilized to calculate p values and compare regressions coefficients, density, and centrality indices. Regression coefficients were presented as connections, while variables represented the nodes in the network.

Results: In comparison to the low GL stratum, the high GL stratum had significantly denser overall ($p = 0.018$) and negative affect network density ($p < 0.001$). The medium GL stratum also showed a directionally similar (in-between high and low GL strata) but

statistically inconclusive association with network density. In contrast to GL, the results of the CT analysis were less conclusive, with increased positive affect density ($p = 0.021$) and overall density ($p = 0.042$) in the high CT stratum compared to the medium CT stratum but not to the low CT stratum. The individual node comparisons across strata of GL and CT yielded only very few significant results, after adjusting for multiple testing.

Conclusions: The present findings demonstrate that the network approach may have some value in understanding the relation between established risk factors for mental disorders (particularly GL) and the dynamic interplay between emotions. The present finding partially replicates an earlier analysis, suggesting it may be instructive to model negative emotional dynamics as a function of genetic influence.

Keywords: emotion dynamics, directed, weighted, network, time-series, genetic, psychopathology, childhood trauma

INTRODUCTION

There is a growing interest in understanding the role of daily-life emotion dynamics underlying psychopathology (van Os et al., 2017). Emotions are considered promising candidates for the study of mechanisms underlying the early expression of subthreshold mental phenomena. From a complex dynamic system theory perspective, alterations in personal emotion dynamics may serve as an early warning sign for a tipping point signaling a transition from a subthreshold state to a clinical state—akin to an electrical signal in epilepsy that is monitored to detect the tipping point before a convulsion (Wichers et al., 2015; Nelson et al., 2017).

In this regard, the network approach provides a useful analytical strategy to gain insight into modeling interactive emotion dynamics, and identifying highly connected emotions that are critical in predicting transition to a more severe state. In recent years, the network approach to psychopathology has brought a novel perspective to conceptualizing mental disorders. Network studies investigate the network of symptoms mutually impacting each other in a variety of mental disorders such as depression and psychotic disorder (Borsboom, 2017). However, one of the primary challenges for the network investigation is that most studies rely on static observations (signs and symptoms) collected from samples with static states (mental disorders) to master a highly fluid phenomenon (Guloksuz et al., 2017).

The experience sampling method (ESM) is designed to prevent recall bias by capturing emotions in real time. ESM uses a rigorous structured diary method for intensive collection of emotions (e.g., sadness, cheerfulness) at random moments during the day, during a certain period (days or weeks), thus providing the essential platform for gathering data for emotion dynamics research (Verhagen et al., 2016).

Recently, the field has advanced to network analysis of ESM data (Pe et al., 2015; Bringmann et al., 2016; Klippel et al., 2017). Emotions have been found to interact with each other

in the network, in which momentarily assessed emotions are represented by a node and the predictive regressive association of that emotion at moment $t–1$ on the same or another emotion at the subsequent moment t, is represented by an edge (Borsboom and Cramer, 2013; Schmittmann et al., 2013). Previous studies demonstrated that an increase in connectivity between affective states was associated with an increased risk for mental disorders (Wichers et al., 2016). Utilizing this approach, the persistence of an emotion over time—inertia—was found to be associated with both current and future depressive episodes (Wichers et al., 2016). By analyzing the auto-regressive coefficient of the emotion, inertia can be studied applying the time series network approach (Kuppens et al., 2012; Bringmann, 2016; Bringmann et al., 2016).

There is growing evidence that the impact of environmental exposure spreads through the symptom network and increase the level of admixture rather than impacting on a symptom domain (Smeets et al., 2012; van Nierop et al., 2014; Guloksuz et al., 2015, 2016). Using data from the general population, previous network investigations showed that the associations between symptoms dimensions and network density increased as a function of the level of environmental exposure (Isvoranu et al., 2016). In a similar fashion, there is some evidence that familial vulnerability operates on increasing connections between symptoms, which in turn leads to a more static and persistent clinical state (Smeets et al., 2014). Given these findings, we previously investigated the network structure of emotional dynamics across environmental and genetic vulnerability strata in a female-female twin population (Hasmi et al., submitted). Although, some differences were observed in the network structure between groups that might be suggestive of an increase in connectivity as a function of vulnerability, findings in general were inconclusive. We now have collected a second large twin sample which can serve as a replication of the previous study in analyzing the impact of vulnerability on emotion dynamics (Hasmi et al., submitted). The present study therefore investigated in a general population mixed-gender twin sample whether genetic liability to psychopathology and childhood trauma (hereafter referred to as "GL" and "CT", respectively) are associated with the network structure of individual emotions—"cheerful," "insecure,"

Abbreviations: CT, Childhood Trauma; GL, genetic liability; EFPTS, East Flanders Prospective Twin Study; ESM, Experience Sampling Method; NA, Negative Affect; PA, Positive Affect; SCL-90-R, Symptom Checklist-90-R; SD, Standard Deviation; MZ, Monozygotic; DZ, Dizygotic.

"relaxed," "anxious," "irritated," and "down"—collected using the ESM method.

In summary, our contributions in the present article are as follows:

- The current study is the first to attempt to replicate a previous network study investigating the question whether genetic factors and exposure to childhood trauma are associated with an alteration in inner psychological functioning at the level of emotions in daily life using the ESM method. Our focus on the emotional dynamics is based on evidence that these dimensions are sensitive to stress exposure and are altered in numerous stress-related mental disorders. It is hypothesized that the higher the levels of risk, the more the emotional network is connected, particularly as regards the negative emotions.

- We utilize a statistical method that is suited to extract and compare network structural parameters from ESM data at the population level dealing with computer tractability matters due to big samples with complex regression models. We inferred network models from the data by assimilating the predictive value of one variable on the other in the time after, to the average connection strengths between our variables of interest.

- We attempt to explain how the impact of genes and early environmental exposure on emotional networks could further be illustrated by examining the moderating effects of genetic risk for psychopathology and childhood trauma on the connections in the network and on three network structural indices. To better evaluate the validity of this approach, we replicate the method in this paper for visualizing the effect of genes and trauma on the network first by displaying visually three different emotional networks of subgroups of participants with increasing childhood trauma and in three emotional networks of subgroups with increasing genetic risk for psychopathology, and second by statistically comparing all those network parameters between the subgroup's networks using permutation testing.

METHODS

Participants

The study sample was derived from the East Flanders Prospective Twin Study register, a population-based prospective register, recording all multiple births in Flanders, Belgium, since 1964 (Derom et al., 2013). Zygosity was determined through sequential analysis based on sex, fetal membranes, umbilical cord blood groups, placental alkaline phosphatase, and DNA fingerprints. Individuals who were registered in the EFPTS and who fulfilled the inclusion criteria were invited to participate in the TwinssCan project, a longitudinal study collecting data on adolescents and young adults between the ages of 15 and 35 years, including twins, their siblings, and parents. The TwinssCan project, which started enrollment in April 2010, is a general population based, ongoing longitudinal study (Derom et al., 2013; Pries et al., 2017). Participants were included if they understood the study procedure and were able to provide valid, reliable, and

complete data. All participants gave written informed consent. For participants below the age of 18 years, parent(s) also signed an informed consent form. Participants were excluded if they had a pervasive mental disorder as indicated by caregivers. The local ethics committee (Commissie Medische Ethiek van de Universitaire ziekenhuizen KU Leuven, Nr. B32220107766) approved the study. For the present study, only twins and siblings who completed the ESM protocol were analyzed, leaving 740 participants.

Measurements
Experience Sampling Method (ESM)

Before the start of the study, the ESM procedure was explained to the participants during an initial briefing session, and a practice trial was performed to confirm that participants were able to understand the 7-point Likert scale response format. During these sessions, subjects were also instructed to complete their reports immediately after the beep, thus minimizing memory distortion. At the start of the protocol, participants received a PsyMate, a custom-made electronic medical Personal Digital Assistant with a touch screen, which was designed to emit a beep-signal at random moments within each of ten 90-min intervals between 07.30 a.m. and 10.30 p.m. on 6 consecutive days. The semi-random beep design prevents participants from anticipatory behaviors and has proven superiority in self-reported adherence in a previous study (Jacobs et al., 2005; Verhagen et al., 2016). At each beep-signal, participants were asked to stop their activity and to enter their current thoughts, context (activity, persons present, and location), appraisals of current situation and mood. To assure reliability and validity, as described in detail before (deVries and Delespaul, 1989; Jacobs et al., 2005), the Psymate records the time at which participants completed the assessment. Furthermore, each beep-signal was accompanied by a 15-min window in which the questionnaire was available to the participant. Reports were required to be completed within 15 min of the beep, with the data recorded as missing outside that interval, as previous work has shown that outside this interval, reports are less reliable and, therefore, less valid (Delespaul, 1995). Also, subjects with fewer than 20 reports were excluded from the analysis.

The items collected by ESM consist of around 40 variables indexing thoughts, current context (activity, social context, and location), appraisals of the current situation, and Emotions. Emotion items at each beep were rated by participants on 7-point Likert scales ranging from 1 = "not at all" to 7 = "very." As in the original study, only 6 emotion variables were chosen for analysis, given their maximum within-person time-lagged variability and therefore minimal floor effect, and given their covering of the whole emotional and core affect spectrum (Russell, 2003). This resulted in the selection of the following emotion items: "cheerful" (positive valence, high arousal), "relaxed" (positive valence, low arousal), "irritated" (loading in both the negative and the positive affect dimensions, high arousal), "down" (negative valence, low arousal), "insecure" and "anxious" (negative valence, high arousal; Hasmi et al., submitted).

Childhood Trauma

The variable CT was assessed using the shortened 25 item version of the 70-item Childhood Trauma Questionnaire (Bernstein et al., 1994, 2003). The CTQ-SF is widely used and validated in various languages including Dutch (Bernstein et al., 2003; Thombs et al., 2009). The continuous variable "CT" reflected the total score of the 25 items on the questionnaire. To visualize the effect of CT on the network, the CT variable was recoded into three categories indexing increasing levels of CT total score and, therefore, severity of trauma (tertile groups). The regression coefficients, for the predictive association between the lag and the current emotions, were calculated for each of the three CT strata before being represented graphically as a network and compared (see below).

SCL-90-R

The Symptom Checklist-90-R (SCL-90-R), a reliable and valid self-report instrument for screening a range of symptoms occurring in the past week, was used to index the overall severity of psychopathology (Wigman et al., 2013). The SCL-90-R consists of nine subscales (Somatization, Obsessive-compulsive, Interpersonal-sensitivity, Depression, Anxiety, Hostility, Phobic anxiety, Paranoid Ideation, and Psychoticism), covering the entire range of psychopathology. The SCL-90-R was assessed twice within an interval of 6 months. First, scores were averaged per participant. Consistent with previous analyses (Wigman et al., 2013), a dichotomous measure of SCL-90-R was used in the analyses, based on the arbitrary cut-off point of 75th percentile. The resulting two-level variable ("SCL-severity") reflected the levels of severity of psychopathology (Wigman et al., 2013).

Genetic Liability to Psychopathology

Genetic liability to psychopathology was determined on the basis of the SCL-90 value (i.e., "low" or "high" psychopathology) in the co-twin and zygosity status, consistent with previous work (Kendler et al., 1995; Wichers et al., 2007; Kramer et al., 2012). This procedure resulted in three categories of "genetic liability": participants with co-twins having a low level of psychopathology (the reference category at lowest genetic liability); participants with a dizygotic (DZ) co-twin with a high level of psychopathology (intermediate level of genetic liability for psychopathology) and participants having a monozygotic (MZ) co-twin with a high level of psychopathology (highest level of genetic liability for psychopathology).

Statistical Analysis

All analyses were performed using Stata version 14.0. (StataCorp, College Station, TX, USA). To take into account the hierarchical structure of the data, multilevel (mixed-effects) linear regression models were fitted using the XTMIXED procedure in Stata, considering that level-one units (multiple observations per individual) clustered into level-two units (level of individual twins), that were nested within level-three units (twin pairs).

Associations between *t*-1 Emotional States and Current Emotional States

Time-lagged variables were used as predictors in the multilevel models (Bringmann et al., 2013). Cheerful at time *t* was predicted by (i) "cheerful," (ii) "relaxed," (iii) "irritated," (iv) "insecure," (v) "anxious," and (vi) "down" at *t*−1 (lag 1). All lagged variables were person mean-centered to disentangle within-subject from between-subject effects, which is now the standard procedure in the field of network analyses (Wang and Maxwell, 2015). The same analysis was performed for each of the other emotional states at time point t (dependent variable) in six separate models. Thus, the six affective states variables at *t* were predicted by all six emotion variables at *t*−1. All lagged emotion variables were entered simultaneously in the model, to assess their independent effects. One example of a regression model is:

$$\text{Cheerful}_{ijk} = (B0+e_{ijk}) + B1 * \text{lag cheerful}_{ijk} + B2 * \text{lag insecure}_{ijk} + B3 * \text{lag relaxed}_{ijk} + B4 * \text{lag anxious}_{ijk} + B5 * \text{lag irritated}_{ijk} + B6 * \text{lag down}_{ijk} + (B7+u7_{ijk}) * \text{time}_{ijk;}$$

Where B0 is the intercept, B1–B7 stand for the regression coefficients, the subscript *i* stands for the assessment level, *j* for individuals, *k* for twin pairs and u7$_{ijk}$ for the random slope of time (see next sub section), and time is the beep number over days (1–50).

As the time between lagged and current moment must be contiguous, and all beep moments were in the waking time period of the day, the first beep of the day was excluded in all analyses. Analyses were performed across 3 strata of CT as well as across 3 strata of genetic vulnerability.

Random Slope of Time

A time variable (i.e., beep number, counting from 1 to 50) was included in all regression models since any association in the network can be interpreted only if no systematic trend is present in the data (i.e., the models are controlled for time effects). Because any trend that may be present could differ across participants, a random slope for time was added to the models at the individual level, representing the standard procedure for analysis in network research (Wang and Maxwell, 2015).

The Construction of Emotion Networks

A complete set of analyses in one stratum yielded 36 unstandardized regression coefficients (B). These coefficients were represented in a graph using the following procedure:

A 6-by-6 matrix with the regression coefficients (B) was constructed. The connection thus denotes the extent to which the emotion variable (e.g., cheerful) at time point *t*−1 predicts another emotion item (e.g., relaxed; $\rightarrow B_{cheerful-relaxed}$) at time point *t*, while controlling for all other variables. The elements on the diagonal are the autoregressive effects (self-loops, e.g., $B_{cheerful-cheerful}$).

This procedure was applied across the 3 strata of CT and the 3 strata of GL, separately (in total 6 graphs). Visualization of networks was obtained using R (qgraph package; Epskamp et al., 2012). Moreover, a value higher than the maximum absolute value of the whole set of regression coefficients, in the 3 strata of CT and then in the 3 strata of GL, was assigned to the argument "maximum" in qgraph to scale the connections widths to allow

for a visual comparison across each set of 3 networks (Epskamp et al., 2012).

Assessment of the Network Structure: Density and Node Centrality

In addition to the individual connections in the network, overall measures can contribute to insight into the differences between networks. Density -also called overall connectivity- is the average of the absolute values of all regression coefficients in each of the networks. Following previous literature that examined the vulnerability underlying emotion density specifically at the level of personality dimensions (neuroticism), using time series networks, two parameters were calculated. Negative density is the average of regression coefficient absolute values, that have both the outcome and the predictor as a negative emotion ("anxious," "irritated," "insecure," "down"). Positive density is the average of regression coefficient absolute values, that have both the outcome and the predictor as a positive emotion ("cheerful," "relaxed"; Bringmann et al., 2016).

Centrality analyses allow for the identification of nodes-emotion items- that are more "central" than others in the network. According to the network theory of psychopathology, the greater the value of a node centrality index, the greater the probability for that node to activate other nodes in the network and create a "domino effect" that would activate a sequence of emotions, negative or positive depending on the connections of that node, from where it is difficult for individuals to get out from (Borsboom and Cramer, 2013). Two well-known centrality indices were calculated per network, allowing for a descriptive comparison across the three genetic liability and the three trauma strata: inward strength and outward strength centrality (Opsahl et al., 2010; Epskamp et al., 2012). In-strength of a certain node is the sum of all edges' weights toward it (that node is the outcome variable). The out-strength of a particular node is the sum of all edges' weights going from it (that node is the independent variable) The first will inform on which affect is the more regulated in the daily emotional experience, and the second on which is the more impactful among the six emotions in the daily life experience. Self-loops (e.g., regression weight between e.g., down at $t-1$ and down at t) are counted both in the inward and in the outward strength, taking into account the fact that self-loops are good indicators of emotion inertia, previously described as an indicator of increased vulnerability and decreased psychological flexibility (Hollenstein, 2015; Wichers et al., 2015).

All density and centrality parameters were calculated using Stata 14.0 (StataCorp, College Station, TX, USA). Permutation testing was used to calculate p-values for comparing them across strata (see below).

Permutation Testing

Mixed-effects models should ideally include random slopes for all time-varying predictor variables (and use fully unstructured covariance matrices for the random effects; Barr et al., 2013). This procedure allows for standard errors, and thus p-values, to be correctly estimated. However, the approach is not feasible in the present context, due to the large number of parameters needed, given that the covariance is unstructured (attempts to fit

such models result in convergence problems). Therefore, a single random slope for *time* was included in the model (see above) and in order to obtain valid p-values, the statistical significance of regression coefficients was examined using permutation tests.

Two different types of permutation tests were performed. The first type was used to obtain valid p-values for each regression coefficient (edge). The second type was performed to compare regression coefficients across different strata of GL and CT.

For the first set of permutations, the value of the outcome variable (e.g., "cheerful" at t) was removed from each record of the original data file and reassigned to the same participant in random order in a copy of the original data set. Because assessments were shuffled within participants, the level of clustering within the data described above was unchanged. Refitting the model based on the permuted data then provides estimates of the model coefficients under the null hypothesis of no association. By repeating this process more than 1,000 times, a distribution of the regression coefficients under the null hypothesis was generated. Then, the observed coefficients were compared with the respective null hypothesis distributions to obtain p-values (i.e., the proportion of times that the coefficient in the permuted data was as large as or larger than the observed coefficient; multiplied by two to obtain a two-sided p-value). Given $2 \times 3 \times 6 \times 6$ tests for statistical significance, Simes correction for multiple testing was applied (Simes, 1986). Graphs derived from the analyses are shown both before and after Simes correction for multiple testing (alpha = 0.0224). While main results are the Simes corrected slopes, presentation of the figures with all the slopes prevents conclusions being directly drawn on differences that are merely the result of differences in power related to sample size in subgroups during the calculation of the p-values.

In the second set of permutations, the values of the CT variable were randomly assigned to the participants in another copy of the original data set. Again, regression coefficients in the original data were compared with regression coefficients under the null hypothesis of no difference in regression coefficients between the CT strata. With this procedure, all regression coefficients of the 36 connections (edges) in the network were tested for differences between the CT strata, regardless of the level of significance obtained with the first type of permutation testing. The same procedure was repeated for the different strata of genetic liability. Again, Simes correction for multiple testing was applied for individual edge differences (alpha 0.000462).

The same permutation testing procedure was applied in order to compare density as well as inward and outward strength parameters between the strata. Assuming independence between each index calculation, no multiple testing correction was applied.

RESULTS

Sample Characteristics

GL analyses included 598 participants (230 monozygotic and 368 dizygotic), given that participants without information on their zygosity status, non-twin siblings, and participants without information on psychopathology in the co-twin were excluded.

CT analyses were performed with 688 individuals. Mean age of the participants was 17.6 years (SD 3.7). Forty percent of the total sample was male. The majority was still living with their parents (86%) and went to school (90%). In addition, 28% had a bachelor degree while only 5% had a low level of education.

The average CTQ-SF sum score was 33.8 (SD 8.1). Demographic data and mean levels of ESM items per subgroup of CT and GL are presented in **Table 1**. In general, the mean level of emotions in the third CT strata was significantly higher than in the first and the second strata. "Down" also differed between the second and the first strata. Except for the difference in "relaxed" between the third and the first strata, there were no differences between the GL strata.

Network Graphs

The networks in **Figure 1** represent the associations between momentary emotion items for 3 levels of CT. For the sake of completeness, both the graphs with only edges that remained significant after Simes correction for multiple testing (i.e., with a $p < 0.0224$) and the complete networks are shown. For example, in the high trauma group, "insecure" at time point $t-1$ was negatively associated with "cheerful" at t ($B = -0.06$). Although, this was different from the regression coefficient in the other two trauma groups, significance disappeared after Simes correction. Thus, after Simes correction, none of the connections differed significantly between the strata.

Figure 2 shows network graphical representations for 3 levels of genetic liability. For example, in the high GL group, "insecure" at time point $t-1$ was associated with "insecure" at t ($B = 0.18$; self-loop).

Structural Characteristics of the Networks

PA density and overall density was higher in the high childhood trauma than in the intermediate trauma network (**Table 2**, **Figure 3**), but density did not linearly increase with increasing

TABLE 1 | Descriptives stratified by childhood trauma and genetic liability.

CHILDHOOD TRAUMA

	Low	Medium	High
Number of subjects (number of assessments)	229 (9,241)	258 (10,438)	201 (7,988)
% Females	71%	56%	53%
% Low education	2%	6%	8%
Mean age (SD)	17.8 (3.66)	17.4 (SD = 3.81)	17.6 (SD = 3.81)
Range	15–33	14–34	15–34
Mean trauma total score (SD)	27.2 (1.38)	32.4 (1.68)	43.1 (9.17)
ASSESSMENT LEVEL			
Cheerful mean (SD overall; between; within)	4.99 (1.49; 0.81; 1.26)	4.88 (1.47; 0.80; 1.24)	4.56 (1.63; 0.97; 1.32)[*†]
Insecure mean (SD overall; between; within)	1.64 (1.17; 0.60; 1.01)	1.70 (1.16; 0.65; 0.98)	1.82 (1.25; 0.72; 1.03)[*]
Relaxed mean (SD overall; between; within)	5.2 (1.48; 0.69; 1.31)	5.15 (1.43; 0.69; 1.26)	4.82 (1.53; 0.76; 1.34)[*†]
Anxious mean (SD overall; between; within)	1.4 (0.93; 0.45; 0.82)	1.46 (0.92; 0.48; 0.79)	1.60 (SD = 1.10; 0.63; 0.91)[*†]
Irritated mean (SD overall; between; within)	2.18 (1.62; 0.88; 1.36)	2.19 (1.52; 0.83; 1.29)	2.48 (1.66; 0.97; 1.36)[*†]
Down mean (SD overall; between; within)	1.59 (1.08; 0.55; 0.92)	1.69 (1.11; 0.64; 0.93)[*]	1.91 (1.27; 0.77; 1.02)[*†]

GENETIC LIABILITY

	Low liability	Intermediate liability	High liability
Number of subjects (number of assessments)	452 (18,338)	90 (3,553)	56 (2,314)
% Females	59%	63%	71%
% Low education	6%	6%	10%
Mean age (SD)	17.5 (3.66)	16.4 (1.89)	17.9 (4.02)
Range	14–34	15–22	15–32
ASSESSMENT LEVEL			
Cheerful mean (SD overall; between)	4.85 (1.49; 0.84; 1.24)	4.76 (1.62; 0.91; 1.34)	4.56 (1.56; 0.93; 1.29)
Insecure mean (SD overall; between)	1.70 (1.16; 0.61; 0.99)	1.79 (1.29; 0.78; 1.06)	1.77 (1.24; 0.73; 1.01)
Relaxed mean (SD overall; between)	5.09 (1.45; 0.71; 1.28)	5.08 (1.47; 0.66; 1.32)	4.84 (1.54; 0.80; 1.32)[†]
Anxious mean (SD overall; between)	1.46 (0.93; 0.47; 0.81)	1.58 (1.11; 0.61; 0.95)	1.62 (1.12; 0.75; 0.86)
Irritated mean (SD overall; between)	2.26 (1.57; 0.89; 1.31)	2.40 (1.70; 0.87; 1.46)	2.29 (1.56; 1.03; 1.25)
Down mean (SD overall; between)	1.71 (1.12; 0.63; 0.94)	1.77 (1.21; 0.68; 1.01)	1.92 (1.35; 0.86; 1.05)

SD, Standard deviation; DZ, Dizygotic twins; MZ, Monozygotic twins; Scl-90, Symptoms checklist. [*]The difference in mean with that of the low subgroup is statistically significant. [†]The difference in mean with that of the intermediate (or medium) subgroup is statistically significant.
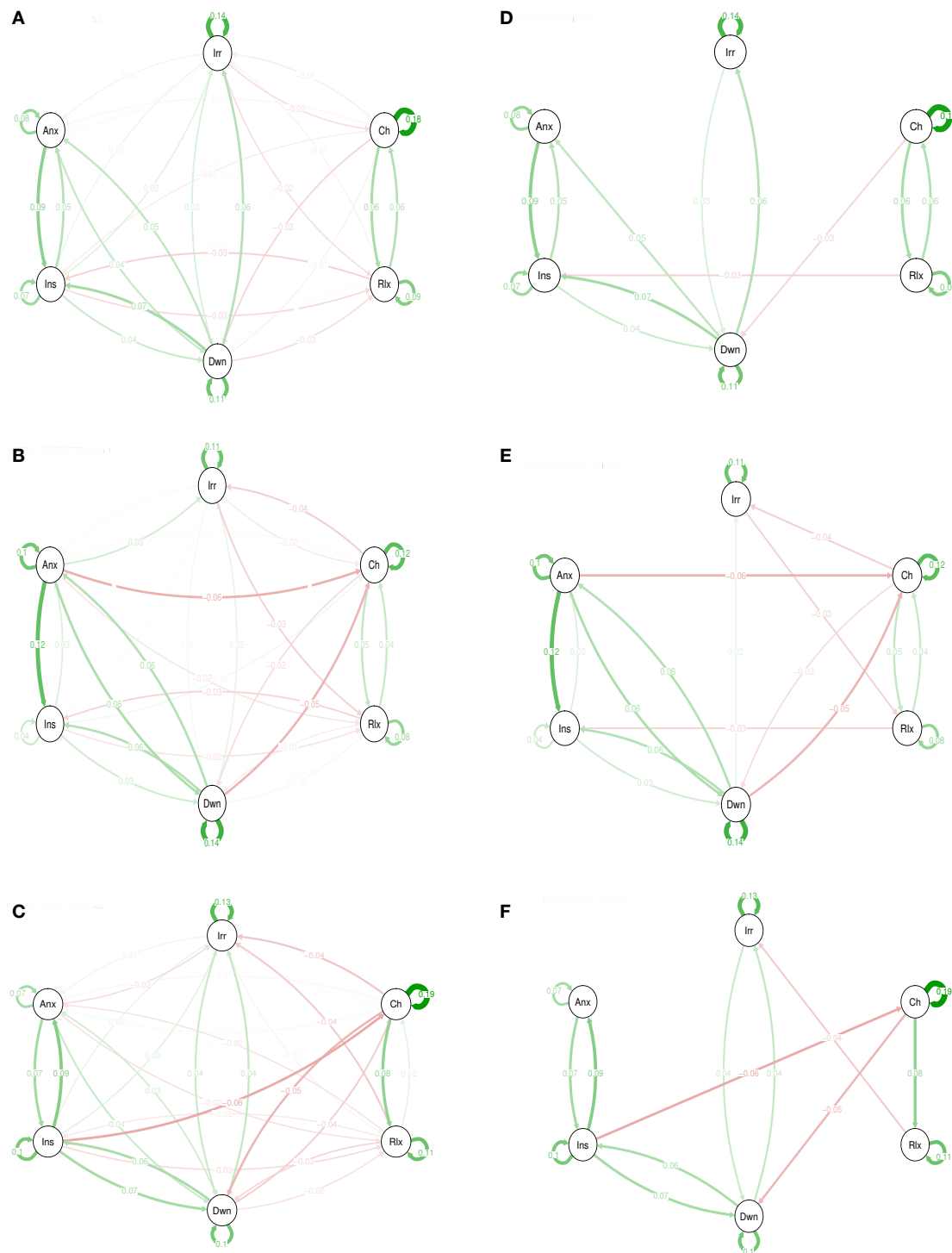
FIGURE 1 | Emotions networks in subjects with low, medium, and high levels of childhood trauma. In this figure, the arrows represent associations over time; i.e., the B coefficient expressing the effect size of the predictive associations. For example, in the low CT network, there is an arrow from "relaxed" to "cheerful," meaning that "relaxed" at $t-1$ predicts "cheerful" at t with a B coefficient of 0.06. Green arrows represent positive associations, and red arrows represent negative associations. The fading of the lines represents the strength of the association and are determined by the regression weights: the more solid the line, the stronger the association (and vice versa). Note that we can predict the emotion item from the previous state of the item itself. These arrows are the self-loops in the network. CT, childhood trauma. Graphs **(A–C)** are for low, medium, and high CT respectively. The Graphs **(D–F)** are for low, medium, and high CT respectively but only with associations that resisted to Simes correction for multiple testing with $p < 0.022$.
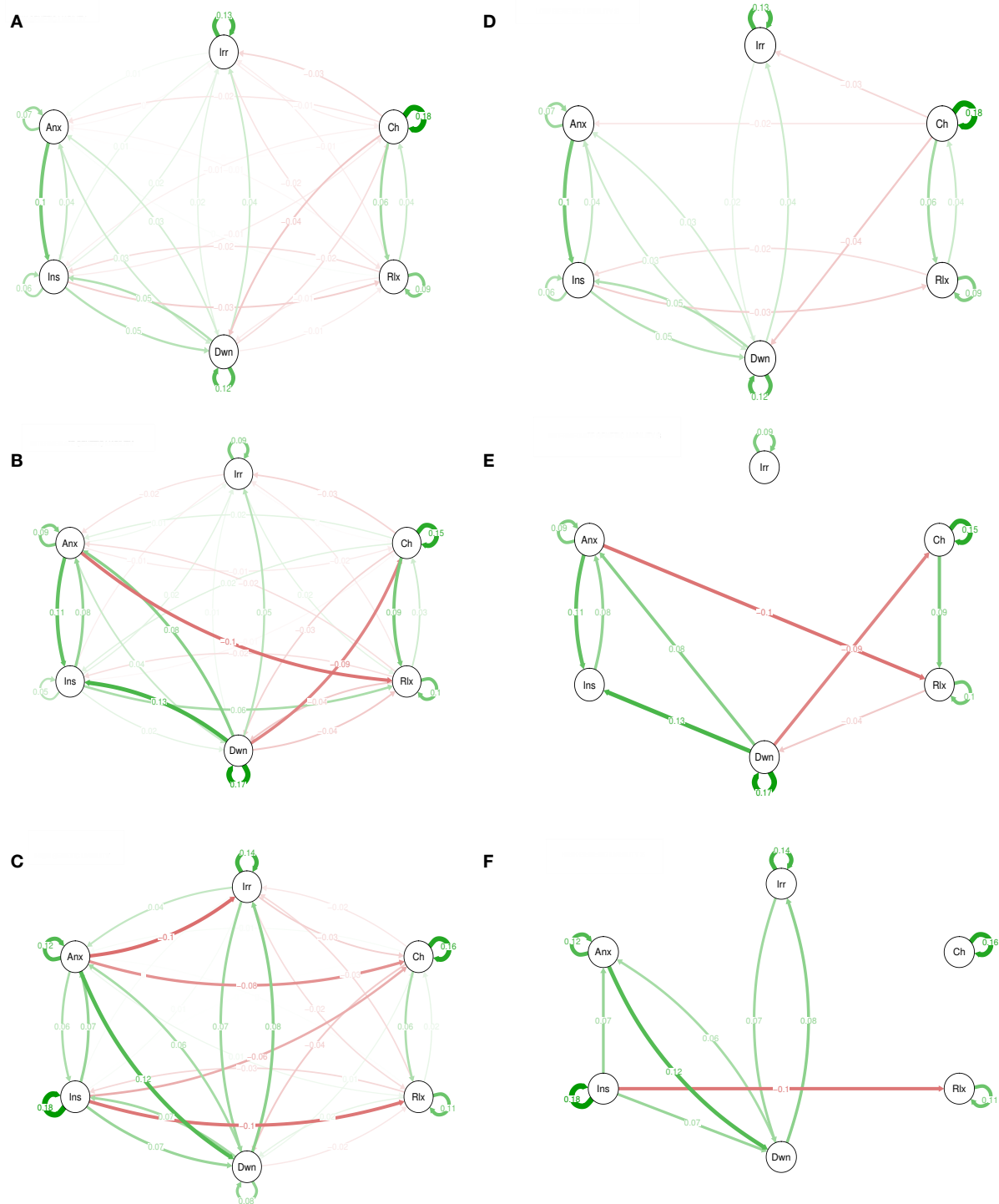
FIGURE 2 | Emotions networks in participants with low **(A)**, intermediate **(B)**, and high genetic liability for psychopathology **(C)**. In this figure, the arrows represent associations over time; i.e., the B coefficient expressing the effect size of the predictive associations. For example, in the low genetic liability network, there is an arrow from "relaxed" to "cheerful," meaning that "relaxed" at $t-1$ predicts "cheerful" at t with a B coefficient of 0.04. Green arrows represent positive associations, and red arrows represent negative associations. The fading of the lines represents the strength of the association and are determined by the regression weights: the more solid the line, the stronger the association (and vice versa). Note that we can predict the emotion item from the previous state of the item itself. These arrows are the self-loops in the network. Graphs **(A–C)** are for low, intermediate, and high GL respectively. The Graphs **(D–F)** are for low, intermediate, and high GL respectively but only with associations that resisted to Simes correction for multiple testing with $p < 0.022$.

TABLE 2 | Emotional density across levels of childhood trauma and genetic liability, respectively.

	Density values			*P*-values of comparison from Permutation tests		
	Low CT	**Medium CT**	**High CT**	**Medium vs. Low CT**	**High vs. Low CT**	**High vs. Medium CT**
PA density	0.09	0.07	0.1	0.06	0.86	0.02*
NA density	0.05	0.05	0.06	0.7	0.6	0.36
Overall density	0.04	0.04	0.05	0.82	0.08	0.04*
	Low gen. liability	**Inter. gen. liability**	**High gen. liability**	**Inter. vs. low GL**	**High vs. low GL**	**High vs. inter. GL**
PA density	0.09	0.09	0.08	0.89	0.66	0.66
NA density	0.05	0.06	0.08	0.22	0.00*	0.09
Overall density	0.04	0.05	0.06	0.05	0.02*	0.61

*$p < 0.05$.

level of trauma. A linear increase was visible in negative and overall density between the GL strata, but only the difference between high and low GL was statistically significant (**Table 2**, **Figure 3**).

"Cheerful" and "down" were most central with respect to outward strength, and cheerful was most central with respect to inward strength across all strata (**Tables 3**, **4**). Centrality of the other emotional items differed between the strata of childhood trauma and the strata of GL, but there was no visible pattern, despite some statistically significant differences. Comparing edges separately, high GL participants showed a significantly stronger "insecure" self-loop than participants with low GL and participants with intermediate GL, but only the difference between high and intermediate GL survived Simes correction (**Table 5**). Only one other connection also survived Simes correction; in the high GL group, "insecure" was followed by a decrease in "relaxed" the next moment (the negative association of "insecure" at $t-1$ with "relaxed" at t; high vs. low; **Table 5**).

DISCUSSION

Using a dynamic network approach, we compared the time-lagged network structures across genetic and environmental risk strata. The primary goal of the study was to identify the impact of CT as an early environmental factor, and GL as a proxy genetic factor, on the structure of a time series network of six emotions—"irritated," "cheerful," "relaxed," "down," "insecure," and "anxious"—at the levels of emotion density, node strength centrality and individual connections (edges). The principal findings were: (i) compared with the low GL stratum, the high GL stratum had significantly denser overall and negative emotion networks, while the medium GL stratum also showed a directionally similar but statistically insignificant association with network density; (ii) in contrast to GL, the results of the CT analysis were essentially inconsistent with our initial hypothesis; (iii) after adjusting for multiple testing, the individual edge comparisons across strata of GL and CT yielded only very few significant results.

Genetic Liability: The Emotion Network Density

Considering the network density across different levels of GL, our current findings suggest an increase in overall and negative density as a function of the extent of GL. As far as we know, differences in density depending on GL have not been studied before. The current study partially replicates an earlier analysis (Hasmi et al., submitted), in which we observed a significant difference in overall density and negative density between high GL and medium GL without a linear increase in density values across the three strata, as opposed to the difference between high GL and low GL with a dose response relation in the present data (Supplementary Material). Also, the individual node comparisons across strata of GL yielded only no significant results (vs. two connections in the present data), after adjusting for multiple testing.

To the degree that higher network density may predict greater symptomatic severity under a high genetic loading, some studies are in apparent agreement with the present results. First, a denser cross-sectional network at baseline was associated with the persistence of clinical depression (van Borkulo et al., 2016; Wichers et al., 2016). Second, in analyses using ESM data, patients with depression, compared to healthy controls, had a higher overall density and negative density, but not a higher positive density (Pe et al., 2015). In agreement with this, higher levels of neuroticism have been associated with a denser emotion network (both overall and negative but not positive; Bringmann et al., 2016). Although, there was no direct estimation of density, several other studies also showed that the more a person shifts toward severe states of psychopathology, the stronger the regression coefficients of mental states at $t-1$ predict mental states at t (Höhn et al., 2013; Wigman et al., 2013). Moreover, according to the results of a recent study that investigated momentary assessed mental states and daily stress while generating three temporal networks in three groups of participants: patients with psychosis, their first-degree relatives, and healthy controls, the number of significant network connections increased in the group of patients with higher familiar risk for psychosis (Klippel et al., 2017). Which is relatively in accordance with the results of the present study; if
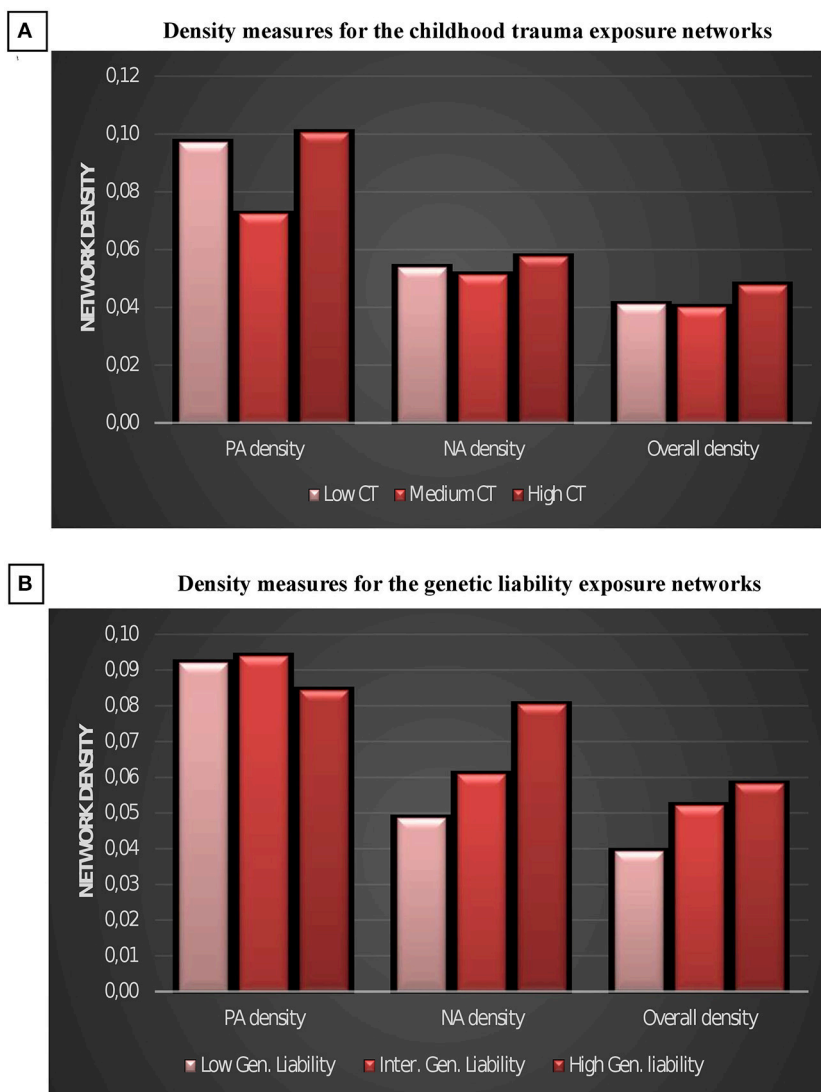
FIGURE 3 | Density measures for the childhood trauma exposure emotions networks **(A)** and for the genetic liability emotions networks **(B)**. CT, childhood trauma; GL, Genetic liability; NA, negative affect; PA, positive affect.

we also consider the fact that, in that work, also connections with non-emotion-related items, e.g., being alone and being active were counted, and that the significance of the connections was not corrected for multiple testing as it is the case in the current work.

Considering the exploratory nature of the time-lagged network analysis of the ESM data and our previous findings, in which we found both higher overall density and higher negative density in the high GL stratum than it was in the medium GL stratum with no difference between low and high GL strata, we err on the side of caution when interpreting the current findings that might be suggestive of an increase in the connectivity of emotions with increasing levels of GL. There might be several explanations for the inconsistency between the previous and the current study. First, consistent with the assumption of the

network theory of psychopathology and with previous work on affect regulation, the expected high between-subject variation might be contributing to reduction in reproducibility (Kuppens et al., 2012). Second, it is plausible to speculate that the differences between characteristics of the two samples may have contributed to inconsistency—the previous study consisted only of female participants with a mean age of 27.7 years. Gender and age differences in terms of symptom profile, vulnerability factors, and epidemiologic features in mental disorders are well identified (van de Water et al., 2016). To the best of our knowledge, there exists no network analysis of ESM data investigating the influence of age and only one study examining a gender effect in a sample of patients with major depressive disorder (MDD) and healthy controls, which showed that women with MDD had a denser negative emotion network than men with MDD, while the gender

TABLE 3 | Node strength centrality across levels of childhood trauma.

	Centrality values			P-values of comparison from Permutation tests		
	Low CT	**Medium CT**	**High CT**	**Medium vs. Low CT**	**High vs. Low CT**	**High vs. Medium CT**
INWARD STRENGTH						
Irritated	0.24	0.20	0.30	0.43	0.32	0.09
Cheerful	0.27	0.31	0.32	0.46	0.36	0.88
Relaxed	0.23	0.20	0.27	0.54	0.52	0.21
Down	0.25	0.29	0.33	0.43	0.11	0.40
Insecure	0.28	0.25	0.27	0.49	0.71	0.77
Anxious	0.20	0.20	0.23	0.97	0.54	0.49
OUTWARD STRENGTH						
Irritated	0.24	0.17	0.21	0.20	0.49	0.54
Cheerful	0.31	0.23	0.37	0.15	0.20	0.01*
Relaxed	0.20	0.17	0.24	0.51	0.44	0.18
Down	0.31	0.34	0.30	0.74	0.88	0.60
Insecure	0.21	0.14	0.37	0.22	0.02*	0.00*
Anxious	0.21	0.39	0.23	0.01*	0.89	0.01*

*p < 0.05.

TABLE 4 | Node strength centrality indices and their relation to genetic liability to psychopathology.

	Centrality values			P-values of comparison from Permutation tests		
	Low gen. liability	**Inter. gen. liability**	**High gen. liability**	**Inter. vs. low GL**	**High vs. low GL**	**High vs. inter. GL**
INWARD STRENGTH						
Irritated	0.23	0.21	0.37	0.34	0.37	0.16
Cheerful	0.27	0.29	0.36	1.19	0.67	0.64
Relaxed	0.22	0.41	0.31	0.01*	0.59	0.16
Down	0.28	0.31	0.40	0.70	0.14	0.37
Insecure	0.25	0.34	0.34	0.24	0.44	0.89
Anxious	0.17	0.31	0.31	0.03*	0.11	1.07
OUTWARD STRENGTH						
Irritated	0.20	0.15	0.31	0.18	0.29	0.05
Cheerful	0.34	0.35	0.28	1.22	0.21	0.36
Relaxed	0.18	0.23	0.20	0.53	1.16	0.52
Down	0.28	0.56	0.32	0.00*	0.91	0.01*
Insecure	0.21	0.23	0.49	1.17	0.02*	0.02*
Anxious	0.21	0.36	0.50	0.25	0.05	0.40

*P < 0.05.

TABLE 5 | Significant edge differences across different levels of GL.

	Differences						Coefficients					
Edges	**Low vs. intermediate GL**		**Low vs. high GL**		**Intermediate vs. high GL**		**Low GL**		**Intermediate GL**		**High GL**	
	Difference	**p**	**Difference**	**p**	**Difference**	**p**	**B**	**p**	**B**	**p**	**B**	**p**
Insecure t-1→ Insecure t,	0.01	00.87	−0.12	00.00*	−0.13	00.01	00.06	00.00[†]	00.05	00.11	00.18	00.00[†]
Insecure t-1→ Relaxed t.	−0.09	00.01	00.06	00.14	0.16	00.00*	−0.03	00.01[†]	00.06	00.05	−0.10	00.01[†]

*P < 0.0004, [†]P < 0.02. Simes corrected alpha for differences across subgroups is 0.0004 and for edge significance is 0.022.

effect was not observed in healthy controls (Pe et al., 2015). In fact, these data—or lack thereof—indicate that there is a need to investigate the impact of basic demographic parameters (e.g., age and gender) on emotion networks before progressing to network analysis of mental disorder constructs in the context of vulnerability.

Childhood Trauma: The Emotion Network Density

Regarding CT, findings were inconsistent, suggesting increased positive density and overall density in the high CT stratum compared to the medium CT stratum but not the low CT stratum, while negative emotion density did not differ across CT strata. In contrast to the current findings, our previous study showed that negative density in the high CT was significantly higher than the medium but not the low CT, with no significant differences in positive and overall emotion density measures across CT strata (see Supplementary Material).

Structural Characteristics of the Networks

Similarly, the individual edge comparisons across strata of GL and CT yielded only very few significant and relatively inconsistent findings after adjusting for multiple testing with Simes correction to avoid spurious conclusions. In the previous female-female twin sample study, statistical comparisons between edges were also inconclusive. Regarding centrality comparisons, only the analysis of the "insecure" node across CT strata yielded a consistent pattern in terms of outward strength, replicating our earlier study. Feeling insecure -also studied as "uncertainty"- was found to be a powerful stressor in previous studies (Greco and Roger, 2003). In previous experiments with replicated results, informing the participants of a low probable electric shock induced more anxiety both at the emotional and physiological level (heart rate and skin conductance) than when the announced probability of the shock was 100% (Lewis, 1966; Epstein and Roupenian, 1970). The replicated finding of "insecure" differences across GL groups may support the notion of a genetic link between "insecure" and negative affect. It may be hypothesized that risk genes impact a brain circuit mediating negative emotion regulation and possibly more specifically emotional reactivity to feeling "insecure."

Strengths and Limitations

The present study replicated the methodology of a recent paper (Hasmi et al., submitted) similar to a series of studies applying network analysis to intensive time series data obtained with ESM to gain insight into dynamic changes in mental states (Bringmann et al., 2013; Klippel et al., 2017). A large number of observations, inherent to the nature of ESM methodology, enabled us to compare three strata of both environmental and genetic exposures. Two other strengths are the use of permutation analyses as a correction for not including all random slopes and the subsequent correction of the alpha for multiple testing for both the p values of the significance of the regression coefficients and the comparison of those regression coefficients individually across networks. Such an approach proved useful,

but it may have negatively affected statistical power and led to type-II error.

There were several limitations. First, only a limited set of momentary emotional mental states was included to overcome convergence problems in the analyses. However, the interest of studying affective mental states was served by this approach. An advantage of analyses in a limited set is also that network graphs are easier to interpret. Second, as data were initially collected from the general population, negative emotion items were relatively rarely reported by participants compared to a clinical population, and, thus, subject to floor effects. This limitation was dealt with by choosing items with the maximum moment-to-moment variation. Third, considering that our participants were young, mainly students, and living with their parents, the results of this study may not be representative of the overall population. Also, the network comparability could be biased by differences in means of emotion items and within-person variances. Means, however, are mostly analogous across GL strata and between the low and medium CT exposure. It hence seems improbable that differences in connection strengths and consequently in network indices, between these latter groups could be attributed to differences in variances. In contrast, the means in the group under high CT exposure are, for most of the emotion items, significantly different from the two other subgroups. Therefore, this could have been part of the reason explaining the lack of replicability between the two studies regarding network density under CT exposure.

Finally, this study is one of few that aimed to compare time series network despite the lack of specific and a valid methodology balancing both type I and type II errors. Further methodological studies are needed; these could, for example, test other advanced methods previously used in comparing cross-sectional networks, and replicate them in ESM based networks across different samples (Fried and Cramer, 2017). Additionally, future work may benefit from a dynamic process approach based on percolation theory (Shang, 2014) and multiplex network models, where interplay between and within several layers (e.g., emotions and vulnerability factors) can be more accurately modeled (Shang, 2015).

CONCLUSION AND FUTURE WORK

The present results represent a partial replication of previous work. The micro-level approach to what could be the phenotypic translation of the genetic liability to psychopathology was demonstrated in both samples, providing a potential link with negative emotion density. The fact that genes impact on the extent to which negative emotions impact each other is important as it helps to expose the complex ways by which genes are affecting mental health. These findings have relevance for future research in psychiatric genetics. First, it may help to explain current problems with replications across studies and, second, it may shine light on the need for novel designs that can take into account the complexity of genetic influence on the development of psychopathology.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01908/full#supplementary-material

REFERENCES

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278 doi: 10.1016/j.jml.2012.11.001

Bernstein, D. P., Fink, L., Handelsman, L., Foote, J., Lovejoy, M., Wenzel, K., et al. (1994). Initial reliability and validity of a new retrospective measure of child abuse and neglect. *Am. J. Psychiatry* 151, 1132–1136. doi: 10.1176/ajp.151.8.1132

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., et al. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse Negl.* 27, 169–190. doi: 10.1016/S0145-2134(02)00541-0

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry* 16, 5–13. doi: 10.1002/wps.20375

Borsboom, D., and Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Bringmann, L. (2016). *Dynamical Networks in Psychology: More than a Pretty Picture?* Ph.D. thesis, Leuven University, KU Leuven.

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., et al. (2016). Assessing temporal emotion dynamics using networks. *Assessment* 23, 425–435. doi: 10.1177/1073191116645909

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., et al. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PLoS ONE* 8:e60188. doi: 10.1371/journal.pone.0060188

Delespaul, P. (1995). *Assessing Schizophrenia in Daily Life: The Experience Sampling Method.* PhD Thesis, Maastricht University Medical Centre.

Derom, C., Thiery, E., Peeters, H., Vlietinck, R., Defoort, P., and Frijns, J. (2013). The East Flanders Prospective Twin Survey (EFPTS). An actual perception. *Twin Res. Hum. Genet.* 16, 58–63. doi: 10.1017/thg.2012.75

deVries, M. W., and Delespaul, P. A. (1989). Time, context, and subjective experiences in schizophrenia. *Schizophr. Bull.* 15:233. doi: 10.1093/schbul/15.2.233

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., and Borsboom, D. (2012). qgraph: network visualizations of relationships in psychometric data. *J. Stat. Softw.* 48, 1–18. doi: 10.18637/jss.v048.i04

Epstein, S., and Roupenian, A. (1970). Heart rate and skin conductance during experimentally induced anxiety: the effect of uncertainty about receiving a noxious stimulus. *J. Pers. Soc. Psychol.* 16, 20. doi: 10.1037/h0029786

Fried, E. I., and Cramer, A. O. J. (2017). Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspect. Psychol. Sci.* doi: 10.1177/1745691617705892. [Epub ahead of print].

Greco, V., and Roger, D. (2003). Uncertainty, stress, and health. *Pers. Indiv. Diff.* 34, 1057–1068. doi: 10.1016/S0191-8869(02)00091-0

Guloksuz, S., Pries, L. K., and van Os, J. (2017). Application of network methods for understanding mental disorders: pitfalls and promise. *Psychol. Med.* doi: 10.1017/S0033291717001350. [Epub ahead of print].

Guloksuz, S., van Nierop, M., Bak, M., de Graaf, R., Ten Have, M., van Dorsselaer, S., et al. (2016). Exposure to environmental factors increases connectivity between symptom domains in the psychopathology network. *BMC Psychiatry* 16:223. doi: 10.1186/s12888-016-0935-1

Guloksuz, S., Van Nierop, M., Lieb, R., van Winkel, R., Wittchen, H. U., and van Os, J. (2015). Evidence that the presence of psychosis in non-psychotic disorder is environment-dependent and mediated by severity of non-psychotic psychopathology. *Psychol. Med.* 45, 2389–2401. doi: 10.1017/S0033291715000380

Höhn, P., Menne-Lothmann, C., Peeters, F., Nicolson, N. A., Jacobs, N., Derom, C., et al. (2013). Moment-to-moment transfer of positive emotions in daily life predicts future course of depression in both general population and patient samples. *PLoS ONE* 8:e75655. doi: 10.1371/journal.pone.0075655

Hollenstein, T. (2015). This time, it's real: affective flexibility, time scales, feedback loops, and the regulation of emotion. *Emot. Rev.* 7, 308–315. doi: 10.1177/1754073915590621

Isvoranu, A. M., Borsboom, D., van Os, J., and Guloksuz, S. (2016). A network approach to environmental impact in psychotic disorder: brief theoretical framework. *Schizophr. Bull.* 42, 870–873. doi: 10.1093/schbul/sbw049

Jacobs, N., Nicolson, N., Derom, C., Delespaul, P., Van Os, J., and Myin-Germeys, I. (2005). Electronic monitoring of salivary cortisol sampling compliance in daily life. *Life Sci.* 76, 2431–2443. doi: 10.1016/j.lfs.2004.10.045

Kendler, K. S., Kessler, R. C., Walters, E. E., MacLean, C., Neale, M. C., Heath, A. C., et al. (1995). Stressful life events, genetic liability, and onset of an episode of major depression in women. *Am. J. Psychiatry* 152, 833–842. doi: 10.1176/ajp.152.6.833

Klippel, A., Viechtbauer, W., Reininghaus, U., Wigman, J., Van Borkulo, C., Myin-Germeys, I., et al. (2017). The cascade of stress: a network approach to explore differential dynamics in populations varying in risk for psychosis. *Schizophr. Bull.* doi: 10.1093/schbul/sbx037. [Epub ahead of print].

Kramer, I. M., Simons, C. J., Myin-Germeys, I., Jacobs, N., Derom, C., Thiery, E., et al. (2012). Evidence that genes for depression impact on the pathway from trauma to psychotic-like symptoms by occasioning emotional dysregulation. *Psychol. Med.* 42, 283–294. doi: 10.1017/S0033291711001474

Kuppens, P., Champagne, D., and Tuerlinckx, F. (2012). The dynamic interplay between appraisal and core affect in daily life. *Front. Psychol.* 3:380. doi: 10.3389/fpsyg.2012.00380

Lewis, P. (1966). Preference behavior in an immediate versus variably delayed shock situation with and without a warning signal. *J. Exp. Psychol.* 72, 847–848.

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. W., and Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder: a review. *JAMA Psychiatry* 74, 528–534. doi: 10.1001/jamapsychiatry.2017.0001

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: generalizing degree and shortest paths. *Soc. Netw.* 32, 245–251. doi: 10.1016/j.socnet.2010.03.006

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., et al. (2015). Emotion-network density in major depressive disorder. *Clin. Psychol. Sci.* 3, 292–300. doi: 10.1177/2167702614540645

Pries, L.-K., Guloksuz, S., Menne-Lothmann, C., Decoster, J., van Winkel, R., Collip, D., et al. (2017). White noise speech illusion and psychosis expression: an experimental investigation of psychosis liability. *PLoS ONE* 12:e0183695. doi: 10.1371/journal.pone.0183695

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110:145. doi: 10.1037/0033-295X.110.1.145

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., and Borsboom, D. (2013). Deconstructing the construct: a network perspective on psychological phenomena. *New Ideas Psychol.* 31, 43–53. doi: 10.1016/j.newideapsych.2011.02.007

Shang, Y. (2014). Unveiling robustness and heterogeneity through percolation triggered by random-link breakdown. *Phys. Rev. E Stat. Nonlin. Soft Matter. Phys.* 90:032820. doi: 10.1103/PhysRevE.90.032820

Shang, Y. (2015). Deffuant model of opinion formation in one-dimensional multiplex networks. *J. Phys. A Math. Theor.* 48:395101. doi: 10.1088/1751-8113/48/39/395101

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751–754. doi: 10.1093/biomet/73.3.751

Smeets, F., Lataster, T., Dominguez, M. D., Hommes, J., Lieb, R., Wittchen, H.-U., et al. (2012). Evidence that onset of psychosis in the population reflects early hallucinatory experiences that through environmental risks and affective dysregulation become complicated by delusions. *Schizophr. Bull.* 38, 531–542 doi: 10.1093/schbul/sbq117

Smeets, F., Lataster, T., Viechtbauer, W., and Delespaul, P. (2014). Evidence that environmental and genetic risks for psychotic disorder may operate by impacting on connections between core symptoms of perceptual alteration and delusional ideation. *Schizophr. Bull.* 41, 687–697. doi: 10.1093/schbul/sbu122

Thombs, B. D., Bernstein, D. P., Lobbestael, J., and Arntz, A. (2009). A validation study of the Dutch Childhood Trauma Questionnaire-Short Form: factor structure, reliability, and known-groups validity. *Child Abuse Negl.* 33, 518–523. doi: 10.1016/j.chiabu.2009.03.001

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B., Waldorp, L., and Schoevers, R. (2016). Association of symptom network structure with the course of longitudinal depression. *JAMA Psychiatry* 73, 412–412. doi: 10.1001/jamapsychiatry.2015.2079

van de Water, T., Suliman, S., and Seedat, S. (2016). Gender and cultural issues in psychiatric nosological classification systems. *CNS Spectr.* 21, 334–340. doi: 10.1017/S1092852916000122

van Nierop, M., Viechtbauer, W., Gunther, N., Van Zelst, C., De Graaf, R., Ten Have, M., et al. (2014). Childhood trauma is associated with a specific admixture of affective, anxiety, and psychosis symptoms cutting across traditional diagnostic boundaries. *Psychol. Med.* 45, 1277–1288. doi: 10.1017/S0033291714002372

van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., et al. (2017). The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. *Depress Anxiety* 34, 481–493. doi: 10.1002/da.22647

Verhagen, S. J. W., Hasmi, L., Drukker, M., van Os, J., and Delespaul, P. A. (2016). Use of the experience sampling method in the context of clinical trials. *Evid. Based Ment. Health* 19, 86–89. doi: 10.1136/ebmental-2016-102418

Wang, L. P., and Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychol. Methods* 20, 63–83. doi: 10.1037/met0000030

Wichers, M., Groot, P. C., Psychosystems, E., and Group, E. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychother. Psychosom.* 85, 114–116. doi: 10.1159/000441458

Wichers, M., Myin-Germeys, I., Jacobs, N., Peeters, F., Kenis, G., Derom, C., et al. (2007). Genetic risk of depression and stress-induced negative affect in daily life. *Br. J. Psychiatry* 191, 218–223. doi: 10.1192/bjp.bp.106.032201

Wichers, M., Wigman, J. T. W., and Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emot. Rev.* 7, 362–367. doi: 10.1177/1754073915590623

Wigman, J. T., Van Os, J., Thiery, E., Derom, C., Collip, D., Jacobs, N., et al. (2013). Psychiatric diagnosis revisited: towards a system of staging and profiling combining nomothetic and idiographic parameters of momentary mental States. *PLoS ONE* 8:e59559. doi: 10.1371/journal.pone.0059559

The Art Gallery Test: A Preliminary Comparison between Traditional Neuropsychological and Ecological VR-Based Tests

Pedro Gamito[1,2]*, Jorge Oliveira[1,2], Daniyal Alghazzawi[3], Habib Fardoun[3], Pedro Rosa[1,2,4], Tatiana Sousa[1], Ines Maia[1], Diogo Morais[1,2], Paulo Lopes[1,2] and Rodrigo Brito[1,2]

[1] Escola de Psicologia e Ciências da Vida, Lusophone University of Humanities and Technologies, Lisbon, Portugal, [2] COPELABS–Cognition and People-Centric Computing Laboratories, Lisbon, Portugal, [3] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, [4] Department of Social and Organizational Psychology, ISCTE – University Institute of Lisbon, Lisbon, Portugal

Ecological validity should be the cornerstone of any assessment of cognitive functioning. For this purpose, we have developed a preliminary study to test the Art Gallery Test (AGT) as an alternative to traditional neuropsychological testing. The AGT involves three visual search subtests displayed in a virtual reality (VR) art gallery, designed to assess visual attention within an ecologically valid setting. To evaluate the relation between AGT and standard neuropsychological assessment scales, data were collected on a normative sample of healthy adults ($n = 30$). The measures consisted of concurrent paper-and-pencil neuropsychological measures [Montreal Cognitive Assessment (MoCA), Frontal Assessment Battery (FAB), and Color Trails Test (CTT)] along with the outcomes from the three subtests of the AGT. The results showed significant correlations between the AGT subtests describing different visual search exercises strategies with global and specific cognitive measures. Comparative visual search was associated with attention and cognitive flexibility (CTT); whereas visual searches involving pictograms correlated with global cognitive function (MoCA).

Keywords: attention, serious games, virtual reality, ecological validity, cognitive assessment

INTRODUCTION

The use of virtual reality (VR) worlds within the field of mental health is a well-established reality. Since the 1990s, VR apps have sprouted up in every area of mental health; today, they cover the most common disorders and disabilities that are identified in the DSM 5 (APA, 2014): Posttraumatic Stress Disorder (PTSD), Obsessive-Compulsive Disorder (OCD), specific phobias and agoraphobias, depression, autism, Alzheimer's disease, and acquired brain disorders. These conditions have all been trialed using VR worlds designed to better immerse patients in the therapeutic process (Brahnam and Jain, 2011).

Virtual reality worlds can reproduce any real-world scenarios – from shopping to taking a ride on the subway. In these worlds, participants can interact freely with the surrounding environment and characters, in the same manner as they do within a real-life environment (Cipresso et al., 2014). By standing in for reality, virtual environments provide ecological validity within a controlled

environment, a combination that is missing from most of the traditional options available to treat mood disorders or to provide exercises for cognitive stimulation and assessment. For a review of the pros and cons of the use of VR applications within the field of mental health and rehabilitation (see Larson et al., 2014).

Another aspect that is also taken into consideration when choosing to use VR apps in treatments is their ability to assess what patients are doing in the virtual world. The system can be programmed to record everything that happens: trajectories, completion times, errors, and indecisions, among other indicators. The VR environment can incorporate in a natural way the actual obstacles that individuals with impairments must resolve in their daily routine. Improvements on indicators (e.g., less errors, lower completion times, less indecisions, and shorter trajectories) may indicate that the cognitive functions required to perform the required tasks were improved. This means that VR apps can assess and provide some insight on behavior performance of the user.

Although it is not possible to ensure their direct translation to a real-world situation, improvements in VR tasks that mimic real life are probably more reliable indicators of actual improvements in performance than are improvements on paper-and-pencil test scores (Silverberg and Millis, 2009; Parsons et al., 2013; Tarnanas et al., 2013; Oliveira et al., 2014; Aubin et al., 2015).

Nevertheless, the degree of functionality of a patient is usually evaluated with self-reports about everyday functioning in basic activities of daily living (ADL: measuring self-care skills), or instrumental activities of daily living (IADL: measuring independent living skills) (Graf and Hartford Institute for Geriatric Nursing, 2008).

Hence, paper-and-pencil tests that are developed under a construct-driven approach, which make them effective in assessing the cognitive constructs of interest, may be ineffective in predicting functional behavior as expressed in everyday tasks because they were not designed to assess actual performance of that behavior (Parsons et al., 2015).

One example of this is the assessment of attention during processes of selecting relevant information from the environment, which have been experimentally assessed through visual search tasks. One of the most used paradigms for this is the comparative visual search paradigm (Pomplun et al., 2001), which consists of a comparison between different shapes across two halves of a display. According to Galpin and Underwood (2005), such tasks involve visual attention for data acquisition, as well as memory to perform the comparison between the two images. Despite the importance of attention and memory for visual search, these tasks comprise the ability to plan and execute an organized pattern of behavior, which also depends on executive functions (Woods et al., 2013; Hardiess and Mallot, 2015).

Our study was built on these notions, as well as on those suggesting that neuropsychological instruments should be adapted to the demands of everyday life activities. We sought to develop a functional visual search task that consisted of an Art Gallery Test (AGT) within a virtual reality setting. This sort of VR-based assessment offers a contextually "realistic"

environment that allows to study "real" psychological and behavioral responses (Bohil et al., 2011).

The AGT was devised to assess the cognitive processes involved in a comparative visual search task while performing matching exercises on paintings that are displayed side-by-side in an art gallery (Rosa et al., 2015, 2016). The three subtests of the AGT represented differences in task difficulty according to the strategy required from participants for visual search (as detailed in the Materials and Methods section).

The aim of the current paper was to study the relation between AGT and standard neuropsychological testing in a sample of adults, testing the hypothesis that performance on the AGT tests is related to the cognitive domains traditionally assessed by visual attention and cognitive functioning assessment scales, and therefore, be a more ecologically valid option compared to traditional neuropsychological testing. For that, the relationships between the AGT subtests with cognitive screening and attention tests were explored. The existence of those correlations can provide evidence to support the proposal of AGT as an alternative to traditional neuropsychological assessment.

MATERIALS AND METHODS

Participants

The sample consisted of 30 university students (25 female) with a mean age of 25 years ($SD = 8.20$), all native Portuguese speakers recruited from a university campus in Lisbon, Portugal. Mean schooling of this sample was 14 years ($SD = 1.9$). Inclusion criteria were: (i) Portuguese native speakers, (ii) university students, and (iii) basic computer skills (mouse). The exclusion criteria were: (i) not having normal or corrected-to-normal vision, and (ii) current or history of psychiatric/neurologic disorders or substance abuse. No participants were excluded due to the exclusion criteria. Moreover, no significant differences were found between genders on age and education ($p > 0.05$).

Measures

The measures used in this study consisted of paper-and-pencil neuropsychological tests and the AGT. The paper-and-pencil tests chosen were two screening tests for general cognitive ability and executive functioning and one specific test to assess attention and executive functions. These paper-and-pencil tests were used as concurrent measures of performance by the AGT. The screening tests are general tests that cover the most important domains of cognitive functioning, whereas the attention test assesses mostly divided and sustained visual attention, as also tested but in a different fashion in the AGT. The first screening test was the Montreal Cognitive Assessment (MoCA) developed by Nasreddine et al. (2005), which is one of the most used screening tests for cognitive impairments. The MoCA has been studied amongst distinct populations from different countries. It has also been validated for the general Portuguese population by Freitas et al. (2011). This test involves the most relevant cognitive domains that contribute to overall cognitive functioning; namely, executive functions, visuospatial abilities, memory, attention, concentration and working memory, language and orientation.

Higher scores in this test reflect better cognitive function (ranges between 0 and 30).

Another goal was to explore the associations of the AGT with executive functions. Thus, the Frontal Assessment Battery (FAB) was also used in the current study. The FAB was developed by Dubois et al. (2000), and has been used ever since as a screening measure of executive functioning. This test assesses six different executive functions: (i) conceptualization, (ii) mental flexibility, (iii) motor programming, (iv) sensitivity to interference, (v) inhibitory control, and (vi) environmental autonomy. Higher scores on the FAB reflect better levels of executive functioning (ranged between 0 and 18).

The Color Trails Test (CTT) is a specific test for divided and sustained visual attention and executive functioning. This test was developed by D'Elia et al. (1996). The CTT consists of an A4 sheet with a number printed in two different colors. The participants in this test are instructed to link the numbered circles in the correct order with a pencil (i.e., increasing numeric order, alternating between colors) as fast as possible and without lifting the pencil from the paper. Two different forms of the CTT were used, which differ in their difficulty. The first trial of the test involves only numbers, whereas in the second trial the participants must shift between numbers and letters. We considered the number of errors (irrespective of the type) and completion time (in seconds). The interference index was also calculated to distinguish between tracking ability in CTT1, which assesses divided attention, and in CTT2 shifting between numbers and letters assesses executive function through cognitive flexibility (Dugbartey et al., 2000). The interference index was calculated according to the following expression:

(CTT2 completion time – CTT1 completion time)/CTT1 completion time)

Higher scores on both these measures are associated with poorer task performance. An additional executive score was also computed from a weighted composite score of the total score of the FAB, errors, and completion times of the CTT2 (weight FAB = 0.5 FAB; weight CTT2 errors = 0.25; weight CTT2 time = 0.25). Thus, the neuropsychological outcomes were global cognition (MoCA total score), executive score (FAB total score and CTT2 completion time), divided attention (CTT1 completion time), and interference index (CTT1 and CTT2 completion times).

The VR test used here was the AGT, developed with Unity 3D 4.6.4. The AGT was originally developed for the Systemic Lisbon Battery 2.0, which comprises several different tasks for cognitive assessment. The AGT consists of three sets of paintings (see **Figures 1**, **2**). In set 1 (subtest A), the participant must spot differences between two paintings that are displayed side-by-side; this was based on the comparative visual search paradigm (Pomplun et al., 2001), requiring the observer to compare the differences between the two halves of the display. Participants were instructed to find seven differences by clicking on the left mouse button. The second and third sets (respectively, subtests B and C) consist of simple visual search tasks with the target stimulus visible during task execution. In set 2 (subtest B), the

participant is required to deconstruct a puzzle according to the details that are displayed on the left side, whereas in set 3 (subtest C), the participant is required to find five details in each painting, that are displayed below the painting; our intention was to create a specific attentional set in their visual search strategy. The design of subtests B and C of the AGT followed the methodology of Baddeley et al. (2001) visual search task using pictograms, but here using pieces of the paintings. Both of these subtests are based on visual search tasks, but subtest B also incorporates planning strategies.

Three different experts with expertise in neuropsychological assessment conceptualized the tasks that comprise the AGT. At this stage, the main goal was to develop a measure that gathers the main aspects of attention ability but with the functionalities of different tasks. The participants took a predefined path while moving from painting to painting (clockwise). The outcomes of this task were based on three different measures: (1) number of mouse clicks during the task (MC), (2) completion time in seconds (CT); and (3) composite scores, which represent the overall performance, by aggregating CT and MC scores that were calculated separately for subtest A, for subtest B, and for subtest C. CTs represent the total duration required to complete each task, since the exercise is displayed and until the participants ends it. It is, therefore, a measure of performance. MC number is an indicator of the number of tries a participant must make to successfully complete the task. A greater number of MCs might indicate that the participant is randomly clicking on the scenario, which represents a guessing-like behavior. In any of the subtests included in the AGT, the number of mouse clicks considered are the ones concerning the task itself, since the VR platform does not register any other clicks besides the ones regarding the task. CSs were the mean scores of MCs and CTs. However, they were applied independently to emulate the original paper-and-pencil tasks. Higher CS scores are supposedly associated with poorer performance.

Procedure

The participants gave their written informed consent to the neuropsychological assessment conducted in this study. The participants were first assessed with the paper-and-pencil tests for neuropsychological assessment. A graduate student supervised by a senior neuropsychologist then assessed the participants with the neuropsychological tests. This assessment was conducted in a soundproofed and dimmed room in the experimental laboratory of the university from which the participants were recruited. After this assessment, the participants headed to another room where they were seated approximately 30 cm from a 17″ TFT monitor of an ASUS i7 CORE laptop computer with 2 GB GEFORCE GT Nvidia graphic board set up with the AGT. The paintings were displayed at the participants' eye level. In the AGT, participants were instructed to look for the paintings and to perform the tasks as quickly as possible. The average time to complete each task was around 10 min. The interaction with the paintings, i.e., spotting the differences (subtest A), deconstructing the puzzle (subtest B), and finding the details (subtest C), was executed through clicking on the left mouse button. The left mouse was also used to move forward, and the right mouse
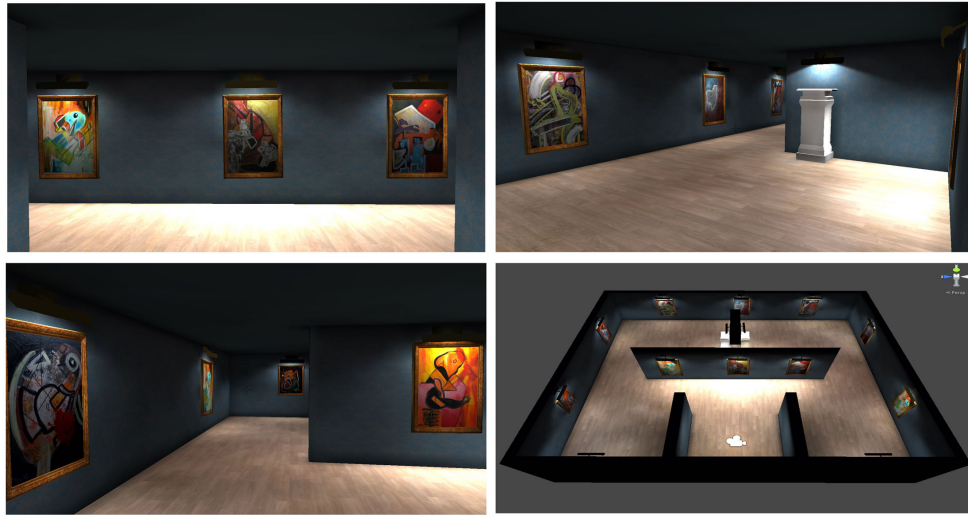
FIGURE 1 | Screenshots from the Art Gallery Test (AGT).
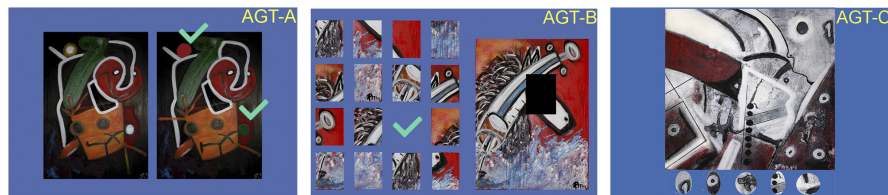
FIGURE 2 | Examples of the AGT's tasks: subtest A (AGT-A); B (AGT-B) and C (AGT-C).

button to move backward. The mouse movement emulated head movements within the virtual environment. Completion times and number of mouse clicks were generated to xls files. The total process (subtests A, B, and C) was concluded in approximately 30 min, after which the participants were dismissed.

Within subtest A, each pair of paintings displayed the same stimuli characteristics, such as luminosity, deepness and color. Within subtests B and C, the pieces/figures of the paintings that were used to deconstruct the puzzle (subtest B) and to be spotted on the paintings (subtest C) were retrieved from the painting that was under search, keeping, in this way, the original properties.

The paintings used are original work of a young Portuguese painter named João Marques (who is also a student at the School of Psychology and Life Sciences of ULHT) who agreed to have his work exhibited in our virtual gallery. Importantly, this ensures that the participants had no prior contact with the art pieces. All the work developed at LabPsiCom is freely available to other researchers on request. AGT has also been tested in touch-screen equipment, such as tablets and smartphones, and is also available on request.

Statistical Analysis

The statistical analysis was performed with SPSS v.21. The first objective was to describe the central tendency, dispersion, and distribution of the neuropsychological and AGT indicators (CT,

number of MC and CS, along with the Z-scores for each of the AGT outcomes). Following this analysis, we computed a composite factor, reflecting the mean number of mouse clicks and completion times, as previously explained. Correlations between these indicators and paper-and-pencil tests results were also calculated, as they can be forthcoming indicators of concurrent validity. This analysis was done for the global cognition (MoCA total score), executive score (FAB total score and CTT2 completion time), divided attention (CTT1 completion time), and interference index (CTT1 and CTT2 completion times). The significant correlations were then explored with a two-step hierarchical linear regression analysis. Given the small sample size, beta effects were estimated using bootstrap sampling with 95% confidence (5,000 samples) to compute robust coefficients (Efron and Tibshirani, 1993), and to investigate the predictive power of the most relevant predictors from the AGT while accounting for age effects.

RESULTS

Descriptive Analysis on Neuropsychological Results

The descriptive analyses of the variables related to the neuropsychological assessment are shown in **Table 1**. The

TABLE 1 | Descriptive analyses of the neuropsychological outcomes.

	M	*SD*±	Skewness	Kurtosis
Montreal Cognitive Assessment				
Visuospatial/executive	3.93	1.02	−0.71	0.50
Naming	2.97	0.18	−5.48	30.00
Attention	4.93	1.34	−0.89	−0.48
Language	2.43	0.68	−0.81	−0.40
Abstraction	1.83	0.38	−1.88	1.66
Delayed recall	2.53	1.53	0.06	−0.84
Orientation	5.93	0.25	−3.66	12.21
Total score	24.37	3.03	−0.19	−0.43
Frontal Assessment Battery				
Conceptualization	2.70	0.47	−0.92	−1.24
Mental flexibility	2.47	0.57	−0.46	−0.75
Motor programming	2.50	0.82	−2.01	4.16
Sensitivity to interference	2.87	0.43	−3.50	12.51
Inhibitory control	2.93	0.25	−3.66	12.21
Environmental autonomy	2.90	0.55	−5.48	30.00
Total score	16.37	1.75	−2.71	10.34
Color Trails Test				
CTT1 errors	0.00	0.00	–	–
CTT1 completion time	47.23	15.17	0.38	1.63
CTT2 errors	0.70	1.18	2.12	5.14
CTT2 completion time	97.70	29.52	0.85	−0.21
Interference Index	1.35	1.65	4.43	21.88

CTT, Color Trails Test.

descriptive statistics were based on the mean scores and dispersion of the data through the standard deviation along with distribution statistics; namely, skewness and kurtosis for the variables from the MoCA, FAB, and CTT. MoCA total score is slightly lower than the normative data for the Portuguese population, according to the age and education levels of our sample (Freitas et al., 2011). Some of the distributions of the neuropsychological variables are negative in asymmetry, which may suggest ceiling effects, particularly in the naming and orientation subtest of the MoCA and all the subtests of the FAB, except for the conceptualization and mental flexibility. The variables concerning the CTT were adjusted to normal distribution apart from the interference index, which was positive in asymmetry with pronounced positive kurtosis.

Descriptive Analysis on the AGT Outcomes

The software generated two different measures (i.e., number of mouse clicks and completion times) for the three subtests. These two indicators were then transformed into composite scores for each subtest of the AGT. Thus, nine different variables were analyzed as outcomes of the AGT: three variables were related to the number of mouse clicks for each subtest; three variables for completion time, and three variables describing the composite scores for subtests A, B, and C. The distribution of these variables is shown through boxplot charts (**Figure 3**).

Bivariate Pearson Correlations of AGT's Indicators with Neuropsychological Outcomes

The AGT's correlations with standard neuropsychological outcomes were computed with the Pearson r coefficient (**Table 2**). The neuropsychological outcomes consisted of global cognition (MoCA total score), executive score (FAB total score, CTT2 errors and completion time), divided attention (CTT1 completion time), and interference index (completion times for CTT1 and CTT2). This analysis showed a correlation between the AGT-A with the interference index from the CTT ($r = 0.452$; $p = 0.012$), and between the AGT-C with global cognition as assessed with the MoCA ($r = −0.420$; $p = 0.021$), suggesting that interference (less cognitive flexibility) is associated with a poorer performance in the AGT-A (i.e., higher scores: more time, more mouse clicks), whereas a poorer performance in the AGT-C is associated with poorer global cognitive function (lower MoCA scores).

Prediction of Cognitive Performance

As considerable evidence has shown that lifespan-associated developmental change is an important covariate of cognitive performance (Rosa et al., 2017), age was included in the first step of the hierarchical regression. In the second step, the composite score of AGT in subtest C was added in the statistical model, since this was the only potential predictor found. The **Table 3** displays the results of the regression analysis. In the first step, age was not a significant predictor of MoCA $B = −0.05$; 95% CI [−0.36, 0.07]; $SE = 0.13$, explaining only 2% of the variance on MoCA. However, when AGT-C total was introduced in the second step, the regression model became significant, accounting for 18% of
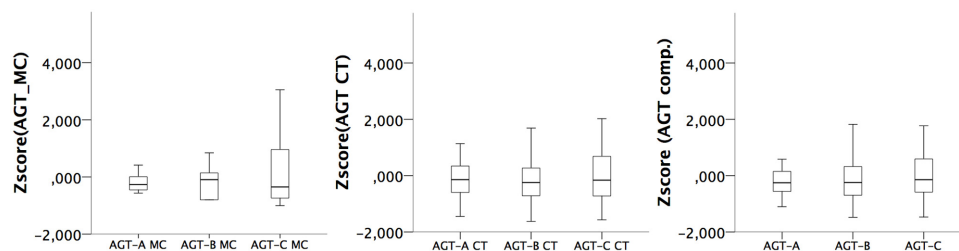
FIGURE 3 | Boxplot charts for AGT-related variables.

TABLE 2 | Bivariate Pearson (r) correlation coefficients between the AGT and neuropsychological outcomes.

	AGT-A total	AGT-B total	AGT-C total
Global cognition	−0.072	−0.138	−0.420*
Executive score	0.220	0.070	0.005
Divided attention	−0.038	0.041	0.013
Interference index	0.452*	0.001	−0.015

*$p < 0.05$. The variables in the first row are related to composite scores for each subtest of the AGT. The variables in the left column are the global cognition retrieved from the Montreal Cognitive Assessment, the executive score calculated from the Frontal Assessment Battery and the Color Trails Test – trial 2, divided attention from the Color Trails Test – trial 1, and the interference index from the Color Trails Test.

the total variance in MOCA. The AGT-C total was a significant predictor of MoCA $B = -0.45$; 95% CI [−0.77, 0.01]; $SE = 0.20$, after controlling for age.

The same analysis was conducted for the AGT-A predicting the interference index, but no significant effects were found in this model. Age was not a significant predictor of the interference index in the first step $B = 0.03$; 95% CI [−0.04, 0.27]; $SE = 0.08$. The second step was significant with the inclusion of AGT-A in the model (21% total variance), although the Beta score for AGT-A remained non-significant, $B = -0.03$; 95% CI [−0.01, 0.08]; $SE = 0.03$.

DISCUSSION

The main goal of this study was to conduct a pilot study to investigate the opportunity to use an ecological neuropsychological test – AGT, as an alternative to traditional paper-and-pencil tests.

Traditional cognitive assessment and rehabilitation materials lack ecological validity and are far from pleasant for patients who have impaired cognitive functions, and need to repeat the same exercise on a daily basis to recover cognitive ability or to minimize the impact of compromised functions (Larson et al., 2014). VR can produce meaningful, ecologically valid, and motivating environments, in which patients can exercise in a gaming fashion (Graf and Hartford Institute for Geriatric Nursing, 2008; Gamito

et al., 2011). The AGT follows these principles, and aims to be an alternative to complement the results from traditional assessment.

To meet this aim, execution scores from a non-clinical population are required. These values represent the average amount of completion time and the average number of attempts (number of mouse clicks) that a healthy participant would take to complete the task at hand. From these results, it would then be possible to find, in future studies with sub-clinical and clinical populations, probable deviations that are likely to occur when a non-healthy participant performs the same task.

The associations between the AGT subtests with the neuropsychological outcomes showed a relationship between subtest A and the interference index of the CTT. These results suggest that the comparative visual search task that was considered more difficult (higher proportion of number of mouse clicks and completion times) was more associated with language-free visual attention processes and cognitive flexibility, as measured in the CTT (Dugbartey et al., 2000), but probably due to the small sample size, the predictive ability of this subtest on the interference index was not significant.

The correlations showed also an association between subtest C, with global cognition as assessed with the MoCA, which was further confirmed using linear regression analysis. The result on subtest C may mean that performance on functional VR tasks is difficult to discriminate through paper-and-pencil tests, as this performance may involve global cognition rather than a specific cognitive function (Oliveira et al., 2017). This was indeed found for the execution of the subtest C of the AGT, which described a visual search task to find the details using pieces of paintings. However, no other associations were found between AGT and executive score, divided attention, nor yet subtest B, which is intriguing given the similarities between subtests B and C. Both tasks were developed using the visual search paradigm, but, considering the visual properties of each of these subtests, the execution strategy may have differed between them. The squared target items in subtest B may have been identified more easily than the small pieces of the paintings in subtest C. A visual inspection of **Figure 3** suggests that mouse clicks were lower on subtest B than on subtest C, which may be indicative of

TABLE 3 | Summary of the Bootstrapped Hierarchical Regression Analysis when predicting global cognition and interference index.

Predictor	Criterion variable	R^2	F	ΔR^2	ΔF	B	SE	CI 95%
Step 1	Global cognition	0.02	0.46 (1, 28)					
Age						−0.05	0.13	[−0.36, 0.07]
Step 2		0.18	2.93 (2, 27)	0.16	5.33* (2, 27)			
Age						−0.02	0.12	[−0.34, 0.09]
AGT-A						−0.44*	0.20	[−0.77, −0.01]
Step 1	Interference Index	0.03	0.82 (1, 28)					
Age						0.03	0.08	[−0.04, 0.27]
Step 2		0.23	3.97 (2, 27)	0.20	6.95* (2, 27)			
Age						−0.04	0.07	[−0.18, 0.13]
AGT-C						−0.03	0.03	[−0.01, 0.08]

*$p < 0.05$. $N = 30$ Values inside the parenthesis are the degrees of freedom. The ΔR^2 and F-values were derived from hierarchical regression analysis. Bootstrap results are based on 5,000 resamples.

different execution strategies to accomplish these tasks. The AGT will require further investigation in clinical samples, with less ceiling effects on neuropsychological data, to better understand the differences between AGT subtests, and whether task difficulty is adjusted to patients with cognitive impairments.

One of the limitations of this study is the small sample. This study was intended as a preliminary approach to the study of AGT as a robust and ecologically valid alternative to traditional neuropsychological assessment. Therefore, the sample size was defined as the minimum necessary to comply with parametric testing based on the central limit theorem. However, the authors will continue this study after proceeding with some technical adjustments and general improvements based on the results of this preliminary study. Another issue concerns gender imbalance, which is justified by the fact that this is a convenience sample from Psychology school where most students are women. Nevertheless, gender was not a factor in the results since the authors previously controlled for its effect. Again, further research with AGT will also take this into consideration.

Although the three AGT subtests were developed to assess attention, the results from this study suggest that the demands of each task and the way they are executed influence the underlying cognitive ability required to accomplish the task (Parsons et al., 2015). Only subtest A correlated with interference index from an attention measure (albeit this effect was not supported in the regression). Subtest C worked as a global measure of cognition, not discriminating what specific cognitive abilities were involved in the execution of this task.

Overall results were more consistent in associating visual search using pictograms (subtest C) with global cognition, although VR comparative visual search (subtest A) may be associated with visual attention and cognitive flexibility as assessed in the CTT.

ETHICS STATEMENT

The study protocol was approved by the EPCV Ethics Committee. All participants gave their written informed consent.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

APA (2014). *DSM 5. Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Arlington, TX: American Psychiatric Association.

Aubin, G., Béliveau, M. F., and Klinger, E. (2015). An exploration of the ecological validity of the virtual action planning-supermarket (VAP-S) with people with schizophrenia. *Neuropsychol. Rehabil.* 28, 1–20. doi: 10.1080/09602011.2015.1074083

Baddeley, A. D., Baddeley, H. A., Bucks, R. S., and Wilcock, G. K. (2001). Attentional control in Alzheimer's disease. *Brain* 124, 1492–1508. doi: 10.1093/brain/124.8.1492

Bohil, C. J., Alicea, B., and Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nat. Rev. Neurosci.* 12, 752–762. doi: 10.1038/nrn3122

Brahnam, S., and Jain, L. (2011). *Advanced Computational Intelligence Paradigms in Healthcare 6: Virtual Reality in Psychotherapy, Rehabilitation, and Assessment. Studies in Computational Intelligence.* Berlin: Springer. doi: 10.1007/978-3-642-16095-0

Cipresso, P., Serino, S., Pedroli, E., and Riva, G. (2014). A virtual reality platform for assessment and rehabilitation of neglect using a kinect. *Stud. Health Technol. Inform.* 196, 66–68. doi: 10.3233/978-1-61499-375-9-66

D'Elia, L., Satz, P., Uchiyama, C., and White, T. (1996). *Color Trails Test Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Dubois, B., Slachevsky, A., Litvan, I., and Pillon, B. (2000). The FAB: a frontal assessment battery at bedside. *Neurology* 55, 1621–1626. doi: 10.1212/WNL.55.11.1621

Dugbartey, A. T., Townes, B. D., and Mahurin, R. K. (2000). Equivalence of the color trails test and trail making test in nonnative english-speakers. *Arch. Clin. Neuropsychol.* 15, 425–431. doi: 10.1093/arclin/15.5.425

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall. doi: 10.1007/978-1-4899-4541-9

Freitas, S., Simões, M. R., Alves, L., and Santana, I. (2011). Montreal cognitive assessment (MoCA): normative study for the portuguese population. *J. Clin. Exp. Neuropsychol.* 33, 989–996. doi: 10.1080/13803395.2011.589374

Galpin, A. J., and Underwood, G. (2005). Eye movements during search and detection in comparative visual search. *Percept. Psychophys.* 67, 1313–1331. doi: 10.3758/BF03193637

Gamito, P., Oliveira, J., Morais, D., Rosa, P., and Saraiva, T. (2011). "Serious games for serious problems: from ludicus to therapeuticus," in *Virtual Reality*, ed. J. J. Kim (Rijeka: InTech Publishing), 527–548. doi: 10.5772/12870

Graf, C., and Hartford Institute for Geriatric Nursing (2008). The Lawton instrumental activities of daily living (IADL) scale. *Medsurg Nurs.* 17, 343–344.

Hardiess, G., and Mallot, H. A. (2015). Allocation of cognitive resources in comparative visual search-individual and task dependent effects. *Vision Res.* 113(Pt A), 71–77. doi: 10.1016/j.visres.2015.05.017

Larson, E. B., Feigon, M., Gagliardo, P., and Dvorkin, A. Y. (2014). Virtual reality and cognitive rehabilitation: a review of current outcome research. *NeuroRehabilitation* 34, 759–772. doi: 10.3233/NRE-141078

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Oliveira, J., Gamito, P., Alghazzawi, D. M., Fardoun, H. M., Rosa, P. J., Sousa, T., et al. (2017). Performance on naturalistic virtual reality tasks depends on global cognitive functioning as assessed via traditional neurocognitive tests. *Appl. Neuropsychol. Adult* 1–7. doi: 10.1080/23279095.2017.1349661 [Epub ahead of print].

Oliveira, J., Gamito, P., Morais, D., Brito, R., Lopes, P., and Norberto, L. (2014). Cognitive assessment of stroke patients with mobile apps: a controlled study. *Stud. Health Technol. Inform.* 199, 103–107. doi: 10.3233/978-1-61499-401-5-103

Parsons, T., McPherson, S., and Interrante, V. (2013). "Enhancing neurocognitive assessment using immersive virtual reality," in *Proceedings of the 17th IEEE Virtual Reality Workshop in Virtual and Augmented Assistive Technology*, Arles, 27–34. doi: 10.1109/VAAT.2013.6786190

Parsons, T. D., Carlew, A. R., Magtoto, J., and Stonecipher, K. (2015). The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical neuropsychology. *Neuropsychol. Rehabil.* 27, 777–807. doi: 10.1080/09602011.2015.1109524

Pomplun, M., Sichelschmidt, L., Wagner, K., Clermont, T., Rickheit, G., and Ritter, H. (2001). Comparative visual search: a difference that makes a difference. *Cogn. Sci.* 25, 3–36. doi: 10.1207/s15516709cog2501_2

Rosa, P. J., Gamito, P., Oliveira, J., Morais, D., Pavlovic, M., and Smyth, O. (2015). "Show me your eyes! The combined use of eye tracking and virtual reality applications for cognitive assessment," in *Proceedings of the 2015 Workshop on ICTs for Improving Patients Rehabilitation Research Techniques*, (New York, NY: ACM), 135–139. doi: 10.1145/2838944.2838977

Rosa, P. J., Gamito, P., Oliveira, J., Morais, D., Pavlovic, M., and Smyth, O. (2016). Uso de eye tracking em realidade virtual não imersiva para avaliação cognitiva [Use of eye tracking in non-immersive virtual reality for cognitive assessment]. *Psicol. Saúde Doenças* 17, 23–31. doi: 10.15309/16psd17 0104

Rosa, P. J., Gamito, P., Oliveira, J., Morais, D., Pavlovic, M., Smyth, O., et al. (2017). Eye movement analysis and cognitive assessment: the use of comparative visual search tasks in a non-immersive VR application. *Methods Inf. Med.* 56, 112–116. doi: 10.3414/ME16-02-0006

Silverberg, N. D., and Millis, S. R. (2009). Impairment versus deficiency in neuropsychological assessment: implications for ecological validity. *J. Int. Neuropsychol. Soc.* 15, 94–102. doi: 10.1017/S1355617708090139

Tarnanas, I., Schlee, W., Tsolaki, M., Müri, R., Mosimann, U., and Nef, T. (2013). Ecological validity of virtual reality daily living activities screening for early dementia: longitudinal study. *JMIR Serious Games* 1:e1. doi: 10.2196/games. 2778

Woods, A. J., Goksun, T., Chatterjee, A., Zelonis, S., Mehta, A., and Smith, S. E. (2013). The development of organized visual search. *Acta Psychol.* 143, 191–199. doi: 10.1016/j.actpsy.2013.03.008

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Computational Psychometrics for the Measurement of Collaborative Problem Solving Skills

Stephen T. Polyak*, Alina A. von Davier and Kurt Peterschmidt

ACTNext, ACT, Inc., Iowa City, IA, United States

This paper describes a psychometrically-based approach to the measurement of collaborative problem solving skills, by mining and classifying behavioral data both in real-time and in post-game analyses. The data were collected from a sample of middle school children who interacted with a game-like, online simulation of collaborative problem solving tasks. In this simulation, a user is required to collaborate with a virtual agent to solve a series of tasks within a first-person maze environment. The tasks were developed following the psychometric principles of Evidence Centered Design (ECD) and are aligned with the Holistic Framework developed by ACT. The analyses presented in this paper are an application of an emerging discipline called computational psychometrics which is growing out of traditional psychometrics and incorporates techniques from educational data mining, machine learning and other computer/cognitive science fields. In the real-time analysis, our aim was to start with limited knowledge of skill mastery, and then demonstrate a form of continuous Bayesian evidence tracing that updates sub-skill level probabilities as new conversation flow event evidence is presented. This is performed using Bayes' rule and conversation item conditional probability tables. The items are polytomous and each response option has been tagged with a skill at a performance level. In our post-game analysis, our goal was to discover unique gameplay profiles by performing a cluster analysis of user's sub-skill performance scores based on their patterns of selected dialog responses.

Keywords: psychometrics, problem-solving, collaboration, clustering, simulation, game, evidence-centered

1. INTRODUCTION

Collaborative problem solving (CPS) is considered as one of the critical skills for academic and career success in the twenty-first century (Griffin et al., 2012). The literature on this topic highlights changing trends that are leading to more employment opportunities that demand collaboration and interaction between people in problem-solving contexts (He et al., 2017; Oliveri et al., 2017). This trend has increased the need in the education industry to address ways to teach and assess these skills (von Davier et al., 2017). In this paper we consider the cognitive and social perspectives of the collaborative problem solving process and examine the circumstances under which collaborative problem solving might best take place to evaluate a participant's level of competency. We outline a structure through which the contributing processes can be monitored and assessed in an electronic environment. In doing so, we reference an emerging discipline called computational psychometrics that is growing out of traditional psychometrics and incorporates techniques from educational

data mining, machine learning and other computer/cognitive science fields. We also introduce our initial work on a collaborative problem solving simulation in which a user is required to collaborate with a virtual agent in order to solve a series of tasks/problems within a first-person maze environment. We demonstrate two techniques based on our knowledge of computational psychometrics:

- Real-time Bayesian evidence tracing that updates sub-skill level probabilities as new evidence is presented.
- A post-game clustering analysis of a user's sub-skill performance scores aimed at defining different profiles of simulation results.

2. MATERIALS AND METHODS

In this section we share our study approach, starting with the identification and selection of the specific CPS sub-skills we monitored. We then describe our simulation/game design, task development and the construction of the conversation tree for the computer agent. Given these constructs, we detail our methods for computational psychometric evidence tagging and continuous evidence tracing. We overview the steps in study execution and data collection. Finally, we define our postgame analysis process that utilizes a set of machine-learning based clustering techniques.

2.1. CPS Sub-skills

For this study, our methodology was to first select a set of collaborative problem solving sub-skills that have been researched and published as part of ACT's investigations into helping people achieve education and workplace success. In "Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success," (Camara et al., 2015) identified facets beyond the well known core academic skills which include the domain-specific knowledge and skills necessary to perform essential tasks in the core content areas of English language arts, mathematics, and science. These additional areas include:

- Cross-cutting capabilities: General knowledge and skills necessary to perform essential tasks across academic content areas. This includes technology and information literacy, **collaborative problem solving**, thinking and metacognition, and studying and learning.
- Behavioral skills: The interpersonal, self-regulatory, and task-related behaviors important for adaptation to, and successful performance in, education and workplace settings.
- Education and career navigation skills: The personal characteristics, processes, and knowledge that influence individuals as they navigate their educational and career paths (e.g., make informed, personally relevant decisions; develop actionable, achievable plans).

As seen above, the cross-cutting capabilities section of the Holistic Framework includes collaborative problem solving as part of a broad, four category enumeration:

1. Technology and Information Literacy

2. **Collaborative Problem Solving**
3. Thinking and Metacognition
4. Studying and Learning

Within the framework, CPS skills are further decomposed into various sub-skills and sub-skill areas. For example, sub-skill areas within CPS include:

- Behavior
- Collaborative Communication
- Problem Analysis
- Solution Planning
- Extended Collaboration (Teamwork)

For this study, we selected 5 sub-skills to gather and analyze for CPS evidence:

- Feature Identification (FI): Identifies the key features of the problem space
- Maintaining a Shared Understanding (MU): Identifying and reconciling gaps in understanding
- Engagement/Interaction (EN): Engagement in the group process and the degree to which that engagement is self-initiated
- Strategy (S): Evidence of establishing a plan of action or policy designed to achieve a major or overall aim
- Evaluate (EV): Recognizing own strengths and weaknesses in relation to others.

2.1.1. Sub-skill Expected Use

Obtaining an assessment at the sub-skill level provides granular evidence to fill in a portion of the Holistic Framework representation of the participant. This diagnostic information can be used to direct toward targeted resources or other remediation steps. It can also be used to provide a representative view of a participant's ability.

2.1.2. CPS Assessments

Society needs assessments that reflect the way people actually teach, learn and work. There are several examples of initiatives and assessments which pioneered a large-scale approach toward measuring CPS skills. These include:

- The Programme for International Student Assessment (PISA) 2015 administered a test of collaborative skills (PISA, OECD, 2013).
- The National Center for Educational Statistics (NCES) commissioned a white paper on the considerations for introduction of CPS in the National Assessment of Educational Progress (NAEP) (NCES, 2017).
- An edited interdisciplinary volume on innovative assessments of collaboration was just published with Springer Verlag (von Davier et al., 2017).
- A special issue of the Journal of Educational Measurement highlighting recent advances in measurement and assessment of cognitive and non-cognitive skills for both individuals and teams, and innoative ways of studying collaboration in education (von Davier, 2017).

- The Smarter Balance Consortium developed an assessment system where performance tasks, including collaborative tasks, are being considered for administration to students as a preparatory experience and are then followed with an individual assessment (Davey et al., 2015).

CPS skills are important for education and career success, but they are difficult to measure. Because CPS is largely enacted as an interactive set of tasks with partners, we need a means to provide a multi-agent setting in which the subjects under assessment can express their abilities. This means providing the opportunity to display the skills in a CPS task for discussion, negotiation, decision making, etc. with another participant, be they a human or simulated agent. In either case, all of these interactive data are referred to as "process data" that offer insight into the interactional dynamics of team members; they are relevant for defining collaborative tasks and for evaluating the results of the collaboration. In the past, these data were not available to scientists at scale. With advances in technology, these complex data can be captured in computerized log files and hence, may allow for meaningful inferences.

The process data from CPS tasks consist of time-stamped sequences of events. From a statistical perspective, these data are time series logs describing the actions and interactions of the users. See Hao et al. (2016) for a discussion of the CPS data. In addition to the process data, if the collaboration is set up in a cognitive (say, math) task, it will also result in outcome data. These types of data are more similar to the outcome data from the traditional tests and indicate if a particular question was answered correctly, or whether the problem was solved (and to what degree it was solved).

Attempting to measure collaboration using a game or other virtual environment is not novel. Neither are the ideas of stealth assessment (Shute et al., 2008b) or evidence centered assessment design (Mislevy et al., 2003; Shute et al., 2008b). However, it is still common to see measurement of collaboration provided by *post hoc* survey data collection (Sánchez and Olivares, 2011; Sung and Hwang, 2013). Measuring through in game data collection techniques holds value, in that more real-time determinations can be made and some of the disadvantages of self-reports (Paulhus et al., 2007) can be avoided, such as self-presentation (Robins and John, 1997).

2.2. Simulation/Game Design

In order to collect data and test hypotheses for this study, ACT developed a CPS game called "Circuit Runner"[1] which allows subjects to play online, in a web browser, with the mission to solve a series of challenges in order to "win" the game. The player needs to collaborate with an automated, virtual agent that has information required to complete the challenges.

In total there are five distinct challenges that range from an agent/player feature discussion around a coded, door-lock panel to a more sophisticated challenge that involves collaborative discovery of a sequence of power transfer steps in order to

succeed. The player navigates from challenge to challenge via a 3-D maze in a first person perspective and is also given continuous access to the agent via a dialog panel which can present prompts and dialog responses from various dialog/conversation trees the player may select. A view of the conversation panel within the game is provided in **Figure 1**. All of the dialog response selections made by the player are recorded in a game "conversation flow" log data file. We can think of the presentation of conversation prompts via the agent as analogous to the presentation of item prompts in a more conventional assessment. The selection of conversation choices by the participant result in item responses captured during the game. Additional telemetry data is gathered including clicks, keystrokes, distance travelled, challenge duration, and dialog selection timing.

2.3. Design Limitations

The canned responses of the dialog tree are a limitation of the current game design. For a more authenticate, natural flow a future design would allow for free text entry or potentially spoken dialog (using speech-to-text to obtain a machine-readable form of that input). Natural language processing (NLP) could then be used to help categorize and ultimately score a given response.

Interaction only occurring between one participant and a virtual agent is another limitation. Allowing multi-human to agent collaboration would add realistic variability and additionally provide another vector of evidence to observe demonstrated CPS skills via human to human interaction.

The game as well as the Holistic Framework's CPS section can be viewed as ultimately stemming from Steiner's proposal regarding group productivity (Steiner, 1972) and a future version that does allow for multi-human interaction would further envelope his proposed performance of a group on a task depended on three factors: the resources available to the group, the requirements of a given task, and the processes by which the group uses to solve a given task. Additionally, the construction of the CPS section of the Holistic Framework looks to Camara et al. (2015) "operationalize the broader construct of collaboration and group work in order to identify specific cognitive skills and strategies that can improve performance." (p. 23)

2.4. Computational Psychometrics

Given these constructs for assessing CPS skills, we consider our methodological basis applying computational psychometrics (von Davier, 2015; von Davier et al., 2017). Computational psychometrics (CP) is defined as a blend of data-driven computer science methods (machine learning and data mining, in particular), stochastic theory, and theory-driven psychometrics in order to measure latent abilities in real-time.

This mixture of disciplines can also be formalized as iterative and adaptive hierarchical algorithms embedded in a theoretical psychometric framework. A similar hierarchical approach to multimodal data was discussed in Khan et al. (2013) and Khan (2015). In a computational psychometrics framework, the test development process and data analysis are rooted in test theory and start with the application of the principle of Evidence Centered Design (ECD) (Mislevy et al., 2006); then, the test is administered as a pilot and the (multimodal) fine grain data are

---

[1]https://cpsgame.stemstudies.com
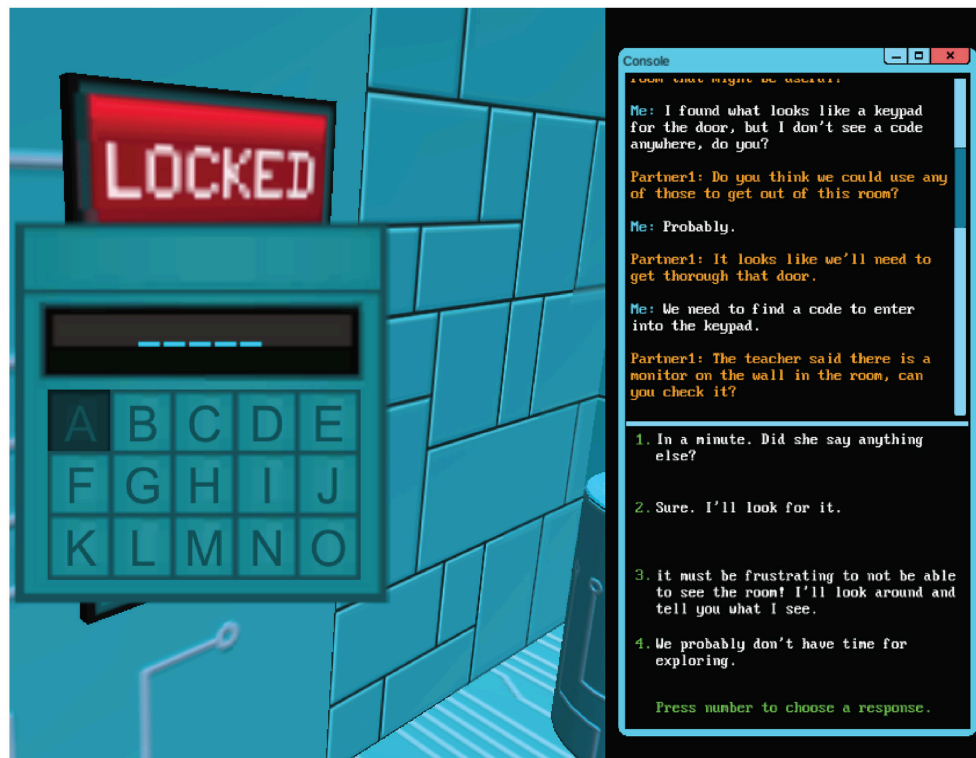
FIGURE 1 | Circuit runner: a CPS dialog panel game screenshot.

collected along with the data from test items (e.g., multiple choice items). This approach is sometimes called a top-down approach because it relies on the expert-based theories. The next step involves a bottom-up approach, in which the data are analyzed by data mining and machine learning algorithms. If new relevant patterns are discovered in the data, these may be incorporated in the revised psychometric models. Next, the psychometric models are revised and the process is repeated with a second round of data collection. One may also apply stochastic processes to the process data. Once the psychometric model is defined and the estimation of the model parameters is stable, the assessment is administered to the population of interest. On the operational data, only supervised machine learning algorithms and already defined and validated psychometric models are further used in order to achieve a stable measurement and classification rules.

This framework involves designing the system (learning and/or assessment) based on theory, identifying constructs associated with the competency of interest, and finding evidence for these constructs from the process data, including video or audio data (Bazaldua et al., 2015). The need for an expansion of the psychometrics framework to include data-driven methods occurred due to the characteristics of the data (dependencies, fine grain size, and sheer volume).

The types of psychometrics models associated with complex data with dependencies have primarily been Bayesian Belief Networks (BBN) (Levy, 2014; Mislevy et al., 2014). BBNs model the probability that a student has mastered a specific knowledge component, conditional on the sequence of responses given to previous elements of a task and eventually to other tasks, whether they are associated with that knowledge component or not (as long as they are part of the network and share at least an indirect connection. BBNs have been applied in games to represent student knowledge and thereby guide the activities of the tutoring system (Corbett and Anderson, 1994; Shute et al., 2008a; VanLehn, 2008; Desmarais and Baker, 2012). BBNs seem attractive for measuring CPS skills, but they have not been adapted to represent the knowledge of multiple individuals simultaneously.

There are stochastic models (point processes, for example) that can be used to model the temporal dynamics of the CPS tasks (von Davier and Halpin, 2013), or hidden Markov models (Soller and Stevens, 2007); there are also models based on the cognitive or social theories such as Agent-based modeling (Bergner et al., 2015) and Markov Decision Process, which is a cognitive model with parameters that describe the goals or beliefs of the agents and which defines behavior as an optimization of expected rewards based on current beliefs about the world (LaMar, 2014). With the aid of data mining techniques we may reduce the dimensionality of the dataset by extracting interpretable patterns which allow research questions to be addressed that would otherwise not be feasible (Romero et al., 2009). This process may help in the scoring process, by assigning different scores to different clusters. Recent papers illustrate

the identification of new evidence to revise the psychometric models (Kerr and Chung, 2012; Kerr, 2015; Zhang et al., 2015).

For the past decade, machine learning algorithms have been used in education to automatically grade written essays; in order to automatically grade and interpret the speech and chat in collaborative interactions we are using similar algorithms; similarly, we can use machine learning for the automatic detection of emotions or affective states during collaboration (Khan, 2015; von Davier et al., 2016).

Bringing the rigor and advanced analysis techniques represented by CP to assess skills that are considered "soft" or otherwise lacking a traditional structure (ie. an acedemic assessment like mathematics), evolve the field. Creating a process to provide for repeatability as well as reducing bias from *ad hoc* methodologies are both desired outcomes of using CP.

In specific practical applications of CP, this hierarchical inference data model may be implemented in simplified or less explicit forms.

2.4.1. Skill Evidence Tagging

For the "Circuit Runner" game, ACT holistic framework researchers designed the tasks and the potential conversation flows, so that they would require participants to collaborate with the virtual agent in a way that would provide evidence of their latent skill ability associated with our selected CPS sub-skills. The dialog tree responses were tagged with one or more sub-skills that were expert judged to provide skill evidence. Furthermore, this evidence was also refined into a level tag using a 3 level enumeration of High, Med, and Low. This was completed by the work of subject matter experts and aligned with the standards set forth in the Holistic Framework (Camara et al., 2015). The experts then verified and approved this tagging and leveling based on committee agreement. The entirety of the tree and the tagging will not be completely disclosed for full reproducibility due to the intellectual property of the dialog tree as well as the proprietary means by which a content expert tagged and developed them.

In **Figure 2** we illustrate this tagging for one item/dialog tree prompt:

> "I am in front of a computer monitor. I have access to the teacher, a map of the maze, and something called an ASCII lookup table. The teacher is talking to me."

and a selected dialog response of:

> "What is the teacher saying?"

This participant event/action presents evidence of CPS skills:

- Monitoring Understanding (MU) at the Med (.2) level (MU.2)
- Engagement (EN) at a High (.3.x) level (EN.3.4)

These items are polytomous and can effectively be scored for a participant based on their sub-skill association and level identification.

2.4.1.1. Bayesian evidence tracing

We can see that conversation flow between the participant and agent provides us with a continuous stream of evidence of a participant's CPS sub-skill, our research question was:

> "Given the real-time, sequential evidence presented via the data of dialog response selections in this game, can we intelligently predict the performance level at each sub-skill?"

The methodology we chose to follow to answer this question used a Bayesian approach related to those typically found in intelligent tutoring systems, such as Bayesian Knowledge Tracing (BKT) (Corbett and Anderson, 1994). The steps to demonstrate this were as follows:

- Extract raw conversation flow game log from a set of played games
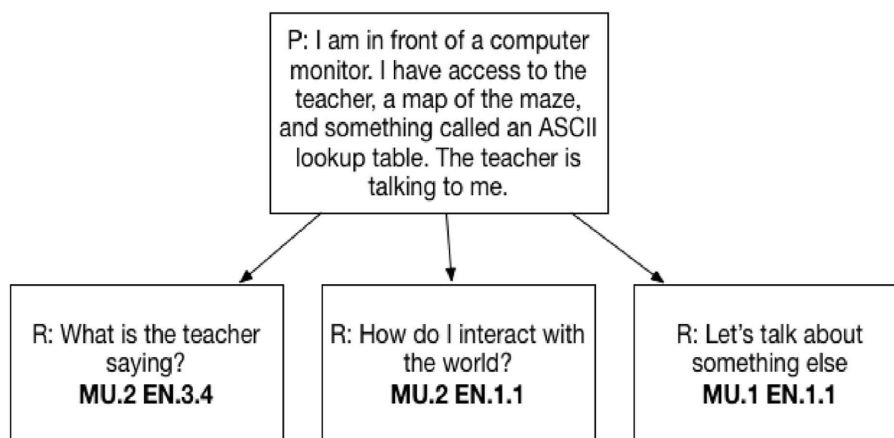
FIGURE 2 | CPS response coding.

- Transform the conversation flow into a flattened file that combines prompt and response and filter out any potential test data
- Generate a 1-Hot encoding of evidence (discussed below)
- Compute Bayesian predictions for all five sub-skills, across each performance level
- Plot the evidence tracing for insight/analysis.

2.4.1.2. Extract

The log data file extracted from the game is outlined in **Table 1**. Each user can have 1 or more sessions and each session can have 1 or more games. In practice though we are typically only interested in 1 game for a single user. As we can see, the log collects the presentation of a dialog tree prompt to the user in a game as row type "P." The prompt presented is recorded in the column "prompt_id." Row type "R" records the response selected by the user in the game for the prompt row immediately preceding it in the log. This raw game log file contained the game session log for several game instances.

2.4.1.3. Transform

Our next step was to flatten this representation so that the prompt and the response rows were combined into a single record. Additionally, we also filtered out data rows that were known to be developer gameplay "user_ids" so that we were only looking at data from actual subject participants. There were also prompt rows followed by some in game action. Instead of a response to that prompt, the user had done something that subsequently caused another prompt to appear. Since there was no response to that initial prompt, it, along with the following action, were also filtered out. Ultimately, $N = 159$ unique games for this analysis.

2.4.1.4. 1-Hot

Taking the flattened prompt/response data, we encoded each game as a single row in a 159x286 matrix. The number of rows is the N count and the number of columns are the three identifiers (session, user, game) plus the 283 potential, selectable dialog responses ($D = 283$). We encoded a "1" if the user selected the identified response at any time during the game. It should be noted that several of the dialog sub-trees can allow a user to loop back through the tree within a single game. If the user selected a particular response more than once in a game we still recorded the selection with a single "1." Otherwise, if the user never selected a particular response during the game the encoding for that column is "0."

2.4.1.5. Compute

Before we introduce our computation of probabilities for the performance levels of a game's CPS sub-skills, let's first review Bayes' theorem and how its application will allow us to trace the evidence over time.

2.4.1.5.1. Bayes theorem. One way to think of Bayes' theorem (Bayes and Price, 1763) is that it gives us a way to update the probability of a hypothesis, H, in light of some body of evidence, E. This way of thinking about Bayes' theorem is called the diachronic interpretation. More precisely, the probability of the hypotheses changes over time as we see new evidence. Rewriting Bayes' theorem with H and E yields

$$p(H|E) = \frac{p(E|H)p(H)}{p(E)} \tag{1}$$

In this interpretation, each term has a name:

- $p(H)$ is the probability of the hypothesis before we see the evidence, called the prior probability, or just "prior."
- $p(H|E)$ is what we want to compute, the probability of the hypothesis after we see the evidence, called the "posterior."
- $p(E|H)$ is the probability of the evidence under the hypothesis, called the likelihood.
- $p(E)$ is the probability of the evidence under any hypothesis, called the normalizing constant.

As an example, let's consider an application of Bayes' Theorem to a simple selection task using two bins to select from. On the performance of this task, we will consider the evidence (E) from a selection event and attempt to compute the probability of two competing hypotheses (H_1) and (H_2). Hypothesis 1 will consider that the selection event happened using bin 1 and hypothesis 2 will consider that the event used bin 2. In **Figure 3** we depict the two bins, bin #1 and bin #2. Bin #1 contains 10 blue widgets (B) and 30 red widgets (R). Bin #2 contains 20 blue widgets (B) and 20 red widgets (R). Let's say that a selection event occurs and the evidence is that of a red widget (R). We will now apply the Bayes' theorem to consider the probability associated with each hypothesis:

1. H_1: The red widget came from bin #1
2. H_2: The red widget came from bin #2

The prior for both $p(H_1)$ and $p(H_2)$ are the same, $\frac{1}{2}$, because we are assuming that red and blue widgets appear equally in each

---

TABLE 1 | Log file format.

Session_id	User_id	Game_id	Time	Type	Prompt_id	Response	...
19	11	1	2015-09-28T15:29:39.302222	P	0.1		
19	11	1	2015-09-28T15:29:49.627254	R		2	
19	11	1	2015-09-28T15:29:49.627254	P	0.3		
19	11	1	2015-09-28T15:29:50.906382	R		2	
...							

bin. The likelihoods are different though, as we can see based on the composition of the bins. Specifically, we have

$$p(E|H_1) = \frac{3}{4} \qquad (2)$$

$$p(E|H_2) = \frac{1}{2} \qquad (3)$$

Putting this all together we can compute the posterior for both hypotheses as:

$$p(H_1|E) = \frac{\frac{1}{2} * \frac{3}{4}}{(\frac{1}{2} * \frac{3}{4}) + (\frac{1}{2} * \frac{1}{2})} = 0.6 \qquad (4)$$

$$p(H_2|E) = \frac{\frac{1}{2} * \frac{1}{2}}{(\frac{1}{2} * \frac{3}{4}) + (\frac{1}{2} * \frac{1}{2})} = 0.4 \qquad (5)$$

We can then state that given the evidence of a red widget we believe there is a 60% chance this was associated with bin #1 and a 40% chance this was associated with bin #2.

2.4.1.6. Response to skill

Given this computation, we can apply it to the evidence and hypotheses we have for the CPS game. In our selection example, the evidence was straight-forward: was the widget blue or red? In the CPS game we need a lookup table for our response to determine which CPS sub-skill and at which performance level the response selection evidence is associated with. The first column of the lookup table combines a prompt identifier and the response, i.e., "0.1–1" (the following row then containing "0.1–2" for the second response of this prompt). The second column contains the noting of skills and levels such as "EN.3.4:FI.2.2:MU.2," that has been tagged by an ACT content expert as providing evidence of:

- Engagement (EN) at a high level (3) (and specifically explanation #4 in that high level)
- Finding Information (FI) at a med level (2) (and specifically explanation #2 in that med level)
- Monitoring Understanding (MU) at a med level (2)

As Mislevy et al. (2014) describe in their application of ECD to interpreting game log data, we can refer to these sub-skills as latent variables, student model variables (SMVs) or

competencies/proficiencies and will denote them using θ, "[the authors] posit that students' performances, characterized by features x_j, arise from some underlying dimensions of knowledge, skill, familiarity, preferences, strategy availabilities, or whatever way we want to characterize them for the purposes at hand. These are called latent variables in the psychometric literature, and student model variables (SMVs), or sometimes competencies or proficiencies, in ECD terminology. We will denote them by θ"

Figure 4 presents a directed graph representation of a multivariate model with parameters that specify conditional distributions of x_j (an instance of a selected CPS dialog response) given θ. The β parameters can represent the "nature and strengths of the relationship" between an x_j and the associated latent variable θ. In this way we can express the relationship between latent variables in our model and the dialog selection evidence using conditional probability tables (CPT) (Mislevy et al., 2014).

2.4.1.7. Conditional probability tables

In our Bayesian example, the $p(H|E)$, or likelihood, was a function of the composition of the bins. In our application of Bayes' rule to the game prediction we will use a conditional probability table for our likelihood term instead. An example of a CPT is shown in **Table 2**. This table was built to provide a modest weighting that indicates a slightly higher likelihood that users will pick responses aligned with their latent variable. Using this table we can explicitly model the type of evidence (high/medium/low performance level, designated by research tagging) which is along
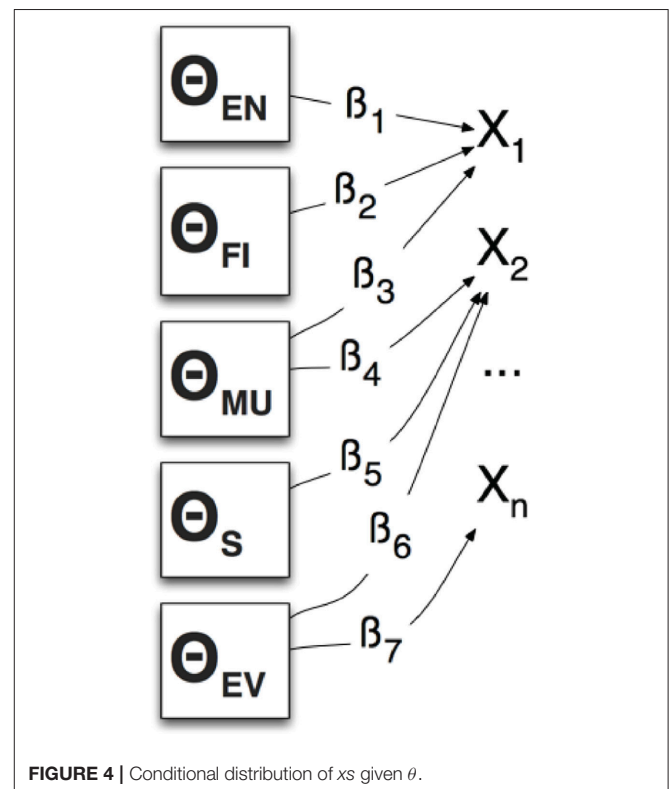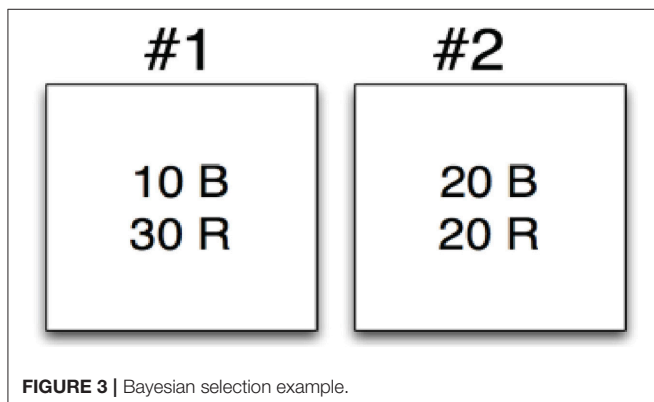
FIGURE 3 | Bayesian selection example.

FIGURE 4 | Conditional distribution of *xs* given θ.

$\theta \mid X_i$	Low	Med	High
θ_{low}	0.4	0.3	0.3
θ_{med}	0.3	0.4	0.3
θ_{high}	0.3	0.3	0.4

the row and the hypothesized performance level of the latent variable (low/medium/high) along the column. Said another way, this table illustrates that if a participant's latent variable is low (row 1) then there is a slightly higher likelihood (.4) that they will select a low tagged response option instead of a medium/med or high level (.3). In practice, there could be a unique CPT created for each item/conversation prompt instance. These unique CPTs might be derived empirically through statistical analysis or could be built using expert judgement. This would allow researchers to fine tune the likelihoods based on the particular item content/difficulty.

2.4.1.8. Evidence tracing

In our Bayesian widget selection example, we presented two possible hypotheses: either the widget came from bin 1 or 2. For the CPS game, we are presented with a response that indicates sub-skill (ss_i) evidence at a particular performance level. As we trace a student's selections we are maintaining three possible hypotheses about the participants latent variable per each sub-skill, viz.

1. Hypothesis: $\theta_{ss_i}^{high}$, Given the evidence to date, the player has a high level for this sub-skill
2. Hypothesis: $\theta_{ss_i}^{med}$, Given the evidence to date, the player has a medium level for this sub-skill
3. Hypothesis: $\theta_{ss_i}^{low}$, Given the evidence to date, the player has a low level for this sub-skill

For each game (G=game_id) then, our algorithm for computing probabilities for the performance levels of a particular sub-skill ss_i is presented in **Figure 5**.

In the initialize step, we set the prior for all hypotheses about a student's sub-skill level at $\frac{1}{3}$, since we have no other evidence. For each dialog response, if it was tagged for the sub-skill then we will recompute the posterior for each hypothesis by incorporating the new evidence. The β value used for the likelihood will be based on a CPT lookup that considers which table is being used for which dialog/response pairing and also what level the skill was tagged with. In our initial application we used the same CPT for all evidence (**Table 2**) but in our future work we intend to work with the dialog content authors to fine tune the application of CPTs based on a more refined judgement of distributions. We demonstrate the results of our tracing in section 3.1.

2.5. Study Execution

We recruited a total of 159 middle school children to play the game. The game was accompanied with a research survey containing personality and background questions. The survey data included age, gender, grades, technology use, and personality

FIGURE 5 | Bayesian evidence tracing algorithm.

facets. This study was reviewed and approved by an independent IRB, the Western Institutional Review Board and carried out in accordance with the Helsinki Declaration. The following steps were approved by the IRB above and were carried out as follows. We recognize 159 samples is rather small and could lead to sample size variation. A second study run is planned with 500 participants and our intention is to perform the same analysis on that data to see how well results concur.

2.5.1. Method of Subject Identification and Recruitment

Prospective subjects were recruited using paper and online advertisements for the study. The recruitment materials encouraged interested parties to visit a secure, publically accessible website with basic information about the study, www.stemstudies.com. Upon arriving at the website, the prospective subjects had the option to complete the consent process and play the game, and then complete the surveys. The game began by asking the visitor for their age to ensure they are eligible to participate.

2.5.2. Process of Consent

This study only collected personally identifiable information as required for fulfillment of the informed consent obligations for the study. This included a username for the child, an email address for the parent, and a name for the parent. Before starting the game, the child created a username, agreed to the informed consent, and provided their parent's email address. After pressing the submit button, the parents were sent an e-mail with background information on the study and a link to the informed consent workflow. Once on the website, the parent will need to provide an email address, first name, last name, and consent to the informed consent document. The informed consent and related materials about the website were available on every page's footer. The parent received a follow-up e-mail 24 h

later informing them that they consented to their child's use of the website. This helped ensure that the parent actually consented to the child's use of the site and allows the parent to revoke consent.

On average, the participants spent around 30 min playing the game. We are currently performing a second run of the study that recruits 500 participants using Amazon Mechanical Turk. In that run we are also including a few more instruments in addition to the game play:

- A pre-survey and post-survey (demographics, background questions).
- A collaborative problem solving questionnaire.
- A situational judgment task assessment involving workplace behaviors relating to collaboration and problem-solving.
- A HEXACO personality assessment. HEXACO is a six-factor structure of personality-descriptive adjectives (Ashton et al., 2004).

2.6. Postgame Analysis

In the postgame analysis, we extracted the raw conversation flow logs from the game and transformed the data to align with the skill/level tagging data provided by the ACT holistic framework researchers. We then used these data to address the following research question:

> "Given the raw data of selected dialog responses across various games played by the students, can we meaningfully group patterns of selections into clusters that may represent different levels of CPS skill evidence?"

Mislevy et al. (2014) demonstrated how traditional assessment approaches relate to emerging techniques for synthesizing the evidence we outlined in our research question. In particular they demonstrate how the models/methods of psychometrics can be leveraged in game-based assessments to collect evidence about aspects of a game player's activities and capabilities.

> "Exploratory data analysis (particularly visualization and hypothesis generation tools) and educational data mining techniques (including recent methods such as unsupervised neural network modeling and …cluster analysis, latent class analysis, and multidimensional scaling) can identify associations among observable features of performance that suggest new student-model variables … Educational data mining is the process of extracting patterns from large data sets to provide insights into instructional practices and student learning. It can often be employed for exactly the tasks of evidence identification: feature extraction based on patterns in data …

Bauckhage and colleagues also discussed the challenges stemming from a similar research question with respect to clustering game behavior data (Bauckhage et al., 2015).

> "the proliferation of behavioral data poses the problem of how to derive insights therefrom. Behavioral data sets can be large, time-dependent and high-dimensional. Clustering offers a way to explore such data and to discover patterns that can reduce the overall complexity of the data. Clustering and other techniques for player profiling and play style analysis have, therefore, become

popular in the nascent field of game analytics. However, the proper use of clustering techniques requires expertise and an understanding of games is essential to evaluate results"

Based on this and other related research (Kerr et al., 2011; Orkin and Roy, 2011; Smith, 2011; Canossa, 2013; Li et al., 2013), it was evident that a machine learning-based, clustering methodology would be useful to explore patterns within our game dialog selection data. In particular we demonstrate an application of game-related, k-means clustering [as reported in other related research (Thurau and Bauckhage, 2010)] vs. other types reported such as Linear Discriminant Analysis (LDA) (Gow et al., 2012) or Mixture Model clustering (Teófilo and Reis, 2013).

2.6.1. Extract
The log data file that is extracted from the game is outlined in **Table 1**. As we can see, the log collects the presentation of a dialog tree prompt to the user in a game as row type "P." The prompt presented is recorded in the column "prompt_id." Row type "R" records the response selected by the user in the game for the prompt row immediately preceding it in the log. This raw game log file contained the game session log for several game instances.

2.6.2. Transform
As we mentioned in our Bayesian workflow, our next step was to flatten this representation so that the prompt and the response rows were combined into a single record. Additionally, we also filtered out data rows that were known to be game developer "user_ids" so that we were only looking at data from actual subjects. There were also prompt rows followed by some in game action. So instead of a response to that prompt, the user had done something that subsequently caused another prompt to appear. Since there was no response to that initial prompt, it, along with the following action, were also filtered out. The N count for this analysis was 159 unique games.

2.6.3. k-means Methodology
The methodology we followed involved these steps:

- Extract raw conversation flow game log from a set of played games
- Transform the conversation flow into a flattened file that combines prompts and responses, and filter out any potential developer gameplay data
- Encode each game as a single row in a 1-Hot encoding of selected dialog responses
- Translate the 1-Hot encoding into 5 datasets corresponding to evidence acquired on all 5 CPS domains
- Perform basic scoring of each game on the 5 CPS domains
- Perform k-means clustering (Steinhaus, 1956) of game domain scores
- Present summary and results of clustering.

2.6.4. Encode/Translate
Taking the flattened prompt/response data we encoded each game as a single row in a 159x286 matrix. The number of rows is the *N* count and the number of columns are the 3 identifiers (session, user, game) plus the 283 potential, selectable dialog

responses ($D = 283$). We encoded a "1" if the user selected the identified response at any time during the game. It should be noted that several of the dialog sub-trees can allow a user to loop back through the tree within a single game. If the user selected a particular response more than once in a game we still recorded the selection with a single "1." Otherwise, if the user never selected a particular response during the game the encoding for that column was "0." Each of the unique dialog prompt/response combinations were coded based on the 5 domains as defined in the CPS game data section

Given this mapping, we were able to create 5 domain evidence matrix variations on the 1-Hot matrix where we substituted the 1,0 with a value of 0,1,2,3 corresponding to the evidence values (no/low/med/high evidence). See **Figure 2**.

2.6.5. Score

Given the 5 domain evidence matrices (as a variation from the 1-Hot encoding) we could then score a game on each of the 5 domains by a simple summing of evidence across each response feature.

$$score^{FI} = \sum_{d=1}^{D} x_d^{FI}$$

$$score^{MU} = \sum_{d=1}^{D} x_d^{MU}$$

$$score^{EN} = \sum_{d=1}^{D} x_d^{EN}$$

$$score^{EV} = \sum_{d=1}^{D} x_d^{EV}$$

$$score^{S} = \sum_{d=1}^{D} x_d^{S}$$

We then reformed the scores into a domain score matrix 159x8 where the rows = N and the columns were the 3 identifiers (session, user, game) plus the 5 summed evidence score for each domain as show in **Table 3**.

2.6.6. Cluster

Using this derived score matrix we then performed an unsupervised learning k-means clustering of the data using the Graphlab-Create library[2]. We selected the K-value based on the following heuristic: $K = \sqrt{N/2.0} = 8$ clusters

2.6.7. K Exploration

Starting with $K = 8$ based on the heuristic value, we continued to evaluate additional potential K-value assignments. The k-means implementation of Graphlab-Create uses the k-means++ algorithm for initial choice of cluster centers. This results in some randomization and variance of cluster assignment with each building of the model. As we visualized the data points with the assignment of the $K = 8$ clusters we noticed similar patterns between several of the clusters. In particular, there appeared to

---

[2]https://turi.com/products/create/

TABLE 3 | Scores matrix.

Session_ id	User_ id	Game_ id	FI_ Score	MU_ Score	EN_ Score	EV_ Score	S_ Score
46	33	211	47	10	26	3	7
57	38	310	39	21	31	4	9
...							

be overlap between 4 sets of 2. This indicated that a 4 cluster assignment may be more appropriate.

We decided to build the model numerous times with a K-value of 8 and compare cluster assignments between these model building runs. We saw that row assignment from the initial cluster assignment didn't always result in classification to the same cluster as on a subsequent build of the model. Sorting the data on the first model build and looking at the cluster classification across the next two builds of the model, we saw some of the same assignments. We subsequently chose $K = 6$ and performed the same multiple run build of the model. Drift was somewhat less, but not significantly so. Setting $K = 4$ and building the model several times showed much less variance in cluster assignment. There was still some drift, but it was significantly less than what we saw with a $K = 8$ and in general cluster assignments persisted across multiple builds of the model even with randomly chosen initial centers.

2.6.8. K-NN Query by Game Id

In addition to the k-means model, we also built a K-Nearest Neighbor (K-NN) model (Arya et al., 1998) using Graphlab-Create which allows us to go back and query the data for games that were similar to a selected game id using a cosine similarity distance metric.

2.6.9. Mixture Model Methodology

There are drawbacks to using the k-means clustering algorithm:

- assumes a specific shape of cluster distributions (spherically symmetric)
- only provides hard assignments to one of the possible clusters.

k-means can be understood as a specific instance of a more generic approach to clustering that is defined by analyzing a mixture of distributions that can be computed using an Expectation Maximization (EM) algorithm (McLachlan and Basford, 1988). Following the same methodology we outlined above to derive our data frame of CPS dialog scores, we re-ran clustering using a mixture of Gaussians approach. This allows us to:

- learn the means and co-variances of each Gaussian distribution (asymmetric, elliptical cluster shapes)
- compute soft assignments to clusters using a Bayesian calculation.

In particular, the EM algorithm works by iteratively running an E-step and M-step where:

1. E-step: estimates cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i | \hat{\mu}_k, \hat{\sum}_k)}{\sum_{j=1}^{K} \hat{\pi} N(x_i | \hat{\mu}_j, \hat{\sum}_j)}$$

2. M-step: maximizes likelihood over parameters given current responsibilities

$$\hat{\pi}_k, \hat{\mu}_k, \hat{\sum}_k | \{\hat{r}_{ik}, x_i\}$$

From a Bayesian perspective, the $\hat{r}_{ik}$ probability represents the responsibility that cluster k claims for observation i expressed as a posterior distribution. This is computed based on $\hat{\pi}_k$, the prior probability of cluster k, and the likelihood that observation i (based on a Gaussian distribution) would be assigned to cluster k given the mean and covariance of the distribution: $N(x_i | \hat{\mu}_k, \hat{\sum}_k)$ divided by the normalizing constant which considers the probability over all possible clusters $\sum_{j=1}^{K} \hat{\pi} N(x_i | \hat{\mu}_j, \hat{\sum}_j)$.

We implemented the code for both the E-step and M-step in Python and ran the implementation over 120 iterations using the MU, FI and EN scores. The S and EV domains were excluded based on their low information content. We also implemented a matplotlib function to plot the computed responsibilities after a specified number of iterations in order to show how the clustering evolved over time. We present those plots in the clustering results section.

3. RESULTS

In the results section, we present visualizations of real-time Bayesian evidence tracing based on a participant's continuous log evidence. We also present the results from our clustering data along with views of cluster data indicators and distributions.

3.1. Bayesian Evidence Tracing Results

Our implementation of the Bayesian algorithm described in **Figure 5** was done in Python using a Jupyter notebook[3] web application. We also used the SFrame API from Graphlab-Create to manipulate the game log data[4]. In order to visualize the sub-skill probabilities over time we initially used matplotlib[5]. An example of the plot for a sample game_id = 114 can be seen in **Figure 6**. This graph shows the increases and decreases of the probability estimates for a participant's EN sub-skill over time. There are three lines because we are tracking each level (high/medium/low) as a separate, but linked variable. All three variables begin using a prior set at .333 and then diverge as the evidence is traced using Bayesian analysis. Additionally we used Tableau[6] to render similar views as can be seen in **Figure 7**. This view allows an analyst to see the predictions of performance levels for each skill, over time, for a single game. The blue area
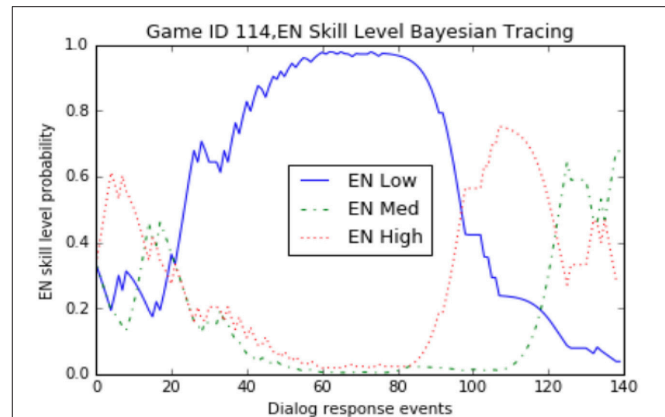
FIGURE 6 | Engagement (EN) sub-skill level probability over time for a single game.

represents a high level, the white area is medium level, and the orange area is the probability of a low level. This view uses an area of fill representation.

Looking at the evidence collected for the single game_id = 114 **Figure 7**, we can see the sub-skills for monitoring understanding (MU) and feature identification (FI) quickly settled on a "medium" level assessment during the first third of the total dialog response interactions. In contrast, the strategy (S) and evaluate (EV) sub-skills settled on a "low" level assessment over the final two thirds of the interactions. The engagement (EN) scores showed fairly dramatic swings between all three performance levels over time, ultimately finishing with a "medium" level assessment. If we were restricted to only looking at the final probabilities (posterior values), we wouldn't have been able to notice these real-time patterns in gameplay. Since the Bayesian Evidence Tracing algorithm is an "anytime algorithm," we are able to directly interrogate this model at any point to determine the current estimate of a user's sub-skill probability.

3.2. Clustering Results

As we described in our methods section, we implemented two clustering approaches, a hard clustering assignment with k-means and a soft clustering assignment using a Gaussian mixture model approach. Additionally, we implemented a K-nearest neighbor (K-NN) mechanism to lookup related games based on the clustering data. The purpose of applying these classification approaches is to look for naturally occuring groups and to determine emergent patterns of game plays or skills.

3.2.1. k-means/K-NN Results
The clustering model using the k-means approach yielded the game counts per cluster as shown in **Figure 8**.

3.2.2. Cluster Characteristics
Now that we have created a clustering model of the game evidence scores, we can inspect the model to see what each cluster might represent about the player/game play evidence of CPS. To that end, we can look at the mean score for each of the 5 domain

FIGURE 7 | Probability (y-axis) over time (x-axis) for a single game (game id = 114) (Blue = High, White = Med, Orange = Low). Engagement (EN), Monitor Understanding (MU), Feature Identification (FI), Evaluation (EV), Strategy (S).

FIGURE 8 | CPS data cluster counts.
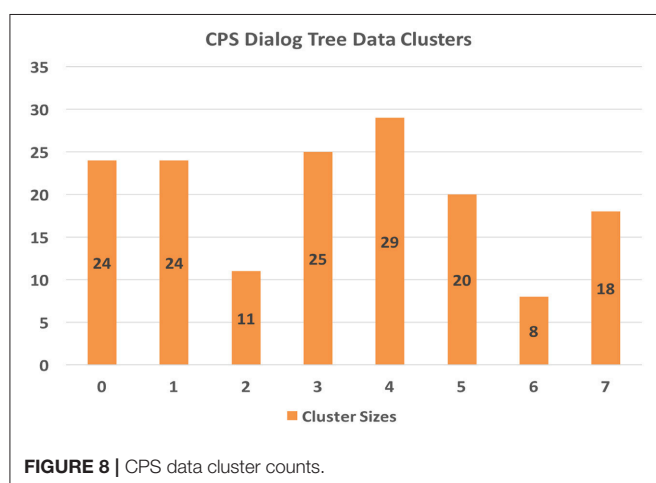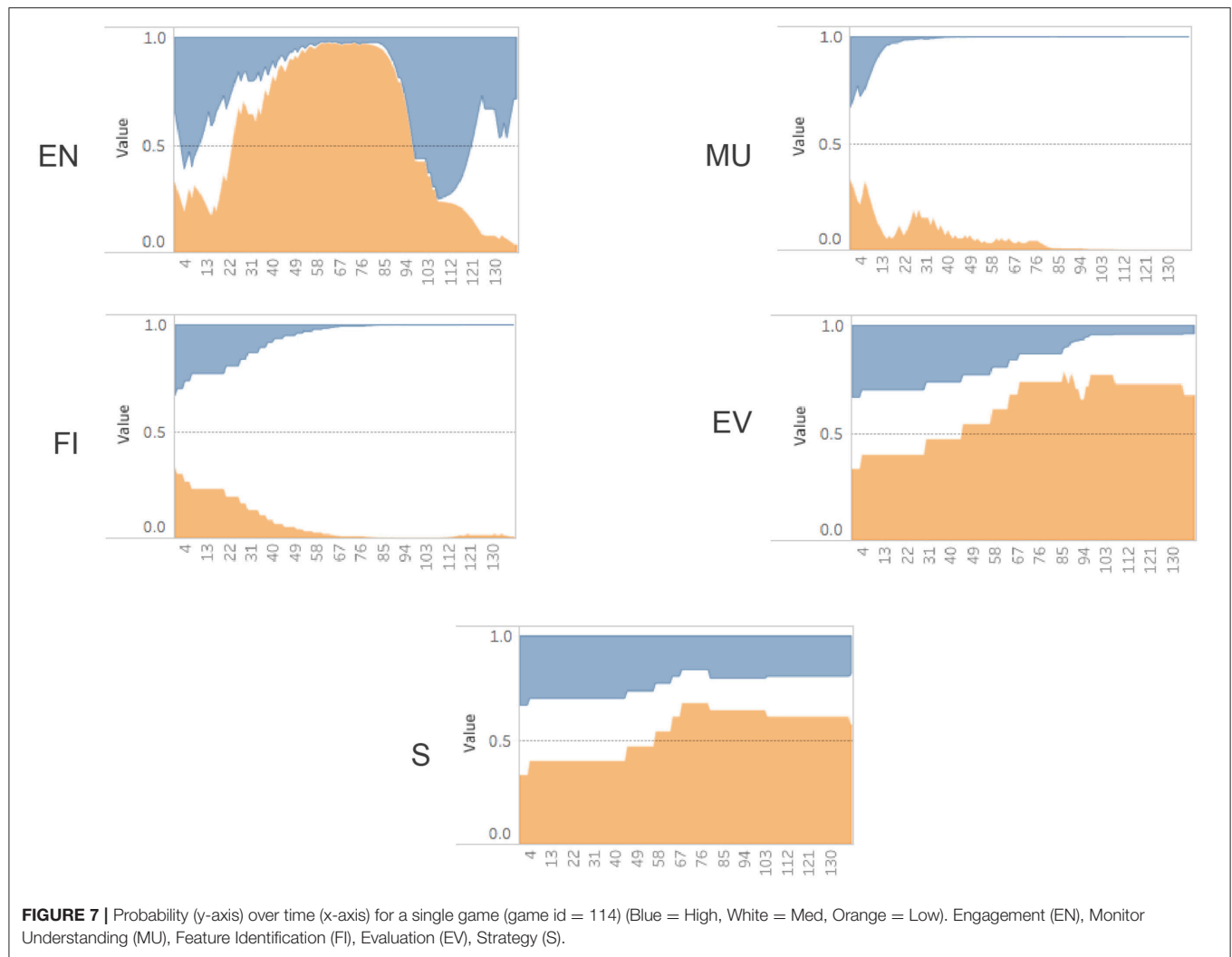
areas for the members of each cluster. The score scales of the 5 domains scores vary considerably, viz. the "EV" and "S" mean scores are much smaller.

For visualization purposes, we normalized the mean scores as follows:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In **Figure 9** we present a graph of the normalized mean scores for each domain across all 8 clusters. We roughly sorted the clusters from left to right within each sub-skill column according to relatively increasing score means. For reference, the raw mean score for each of the sub-skills are: FI = 28, MU = 17, EN = 31, EV = 3, S = 2, and the raw standard deviation for each is: FI = 18.49129825, MU = 10.52478206, EN = 21.48635757, EV = 2.128990584, S = 3.603795468.

Cluster 2 ($N = 11$) represents the games that exhibit the highest CPS scores across nearly all domains (except for FI), whereas cluster 4 ($N = 29$) represents the games that exhibit the lowest CPS scores. Given that we didn't filter out incomplete games, i.e., games where subjects did not make it all the way through the final challenge, it is likely that cluster 4 represents many of these incomplete games. Cluster 6 ($N = 8$) game plays excelled at FI and presented very good scores across the board

FIGURE 9 | CPS data cluster domain score means.

FIGURE 10 | Tableau visualization of scores by cluster. Each color/column is a cluster. A dot represents a game score in CPS sub-skill {EN,FI,MU}. A black line shows a cluster mean score.
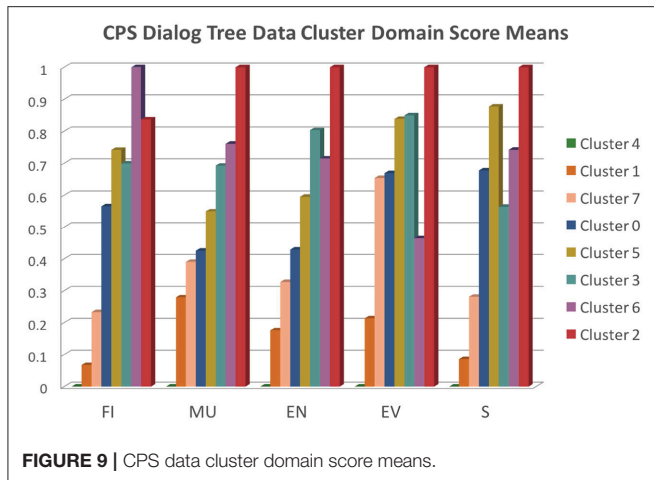
as well. Cluster 3 ($N = 25$) games provided a balanced set of very good scores, especially in EN and EV. Cluster 5 ($N = 20$) game plays excelled at EV and S. Cluster 1 game plays ($N = 24$) provided fairly weak evidence of CPS skills overall, whereas clusters 7 ($N = 18$) and 0 ($N = 24$) presented low to average scores.

We also loaded the data into a Tableau workbook[7] to analyze the cluster characteristics using various worksheets. In that analysis, we saw a vertical distribution of normalized scores grouped by score feature (EN, FI, MU, S, EV) for each of the 8 clusters that showed that while EN, FI, and MU features appeared to have fairly tightly grouped cluster values the features values from S, EV appeared to be much more diffuse within a cluster. As EN, FI, and MU are the important feature drivers of the cluster characteristics we looked at a similar view. That allowed us to examine the cluster distributions across a range of score groupings over EN, FI and MU. In **Figure 10** we re-arrange the data to illustrate the vertical cluster scores (the black line indicates the mean) with each column as a cluster.

3.2.3. K-NN Query by Game Id

In addition to the k-means model, we also built a K-Nearest Neighbor (K-NN) model (Arya et al., 1998) using Graphlab-Create, which allows us to go back and query the data for games that were similar to the source game using a cosine similarity distance metric.

3.2.4. Mixture Model Results

In **Figure 11** we represent how our application of an EM algorithm learned the dialog score cluster responsibilities over a series of iterations. For 2-D visualization purposes we just show the MU/FI features. The color of each dot represents a blending of cluster probabilities.

As we can see the Mixture Model approach updates the cluster distribution shapes over each iteration, effectively learning the mean and covariance of each distribution. In **Figure 12** we plot the final shape of the cluster distributions ($k = 4$), again limiting

---

[7]http://www.tableau.com
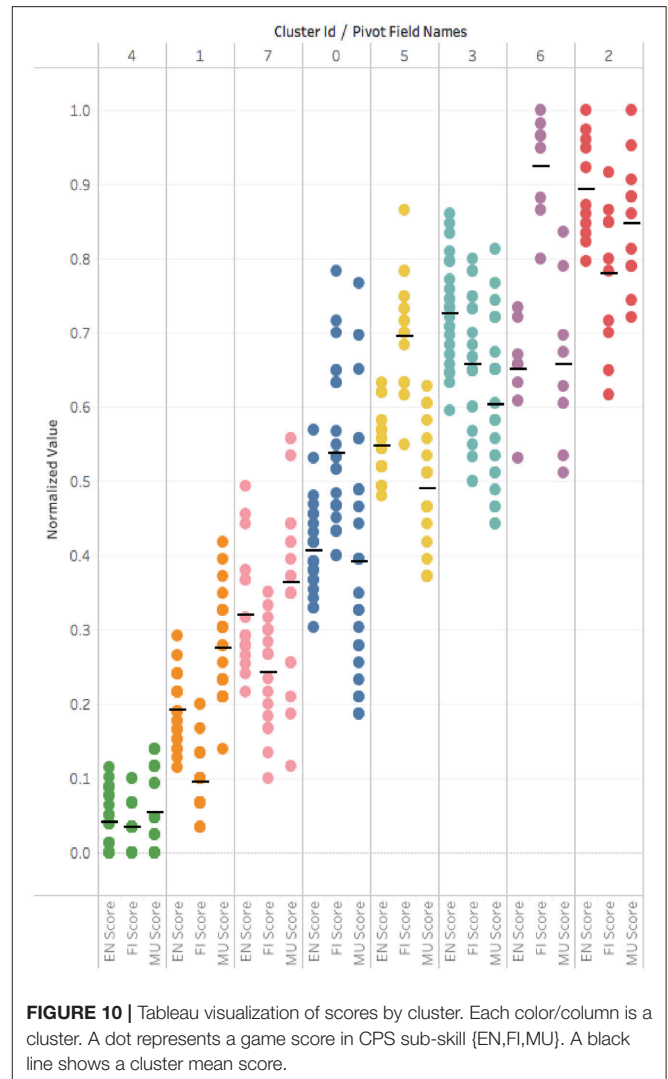
this to just the MU and FI score dimensions. As we can see, this method of clustering allowed the model to learn asymmetric elliptical cluster shapes and also provided us with probabilistic assignments of each observation to any of the clusters. Thus, we are able to represent more robust cluster characterizations beyond a simple in/out hard assignment.

Our interpretation of these data is that the observations in the upper right cluster represent players that were exhaustively exploring the dialog trees which resulted in maximizing their dialog scores. The next cluster to the left represents players who were focused on getting just the data they needed in their collaboration to complete the challenges. The two far left, bottom clusters represent players that were not engaged and probably didn't play through to the final challenge.

4. DISCUSSION

In this paper we have demonstrated the application of computational psychometrics to gathering insights into a
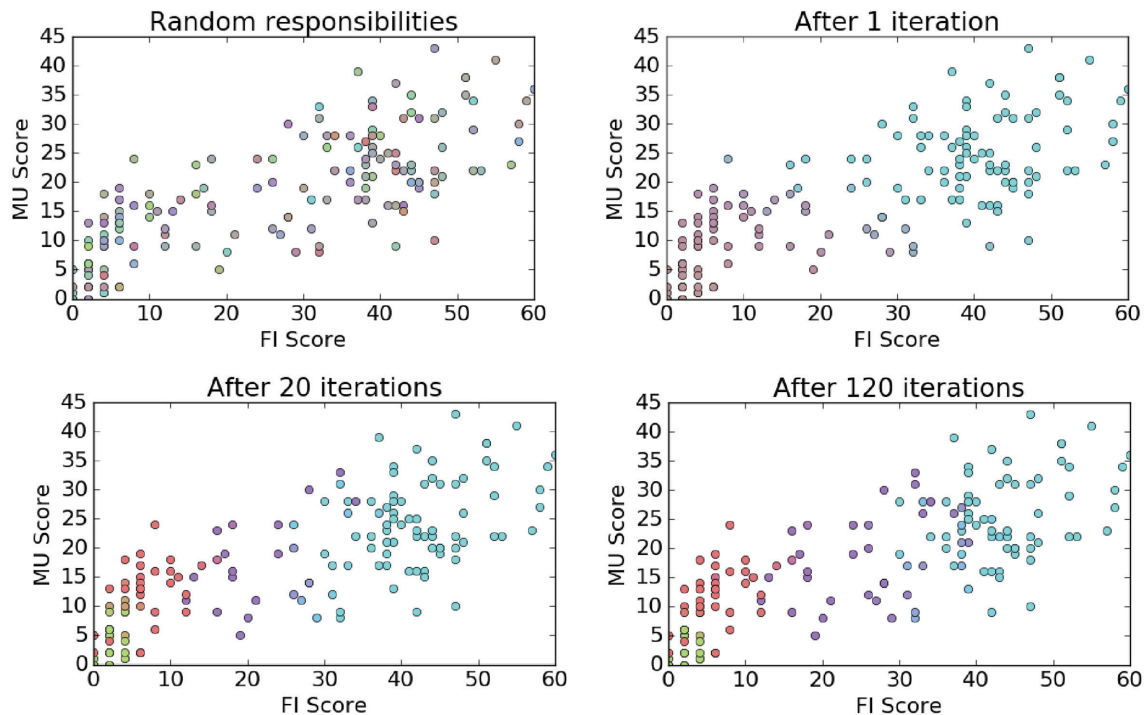
FIGURE 11 | EM clustering visualization.
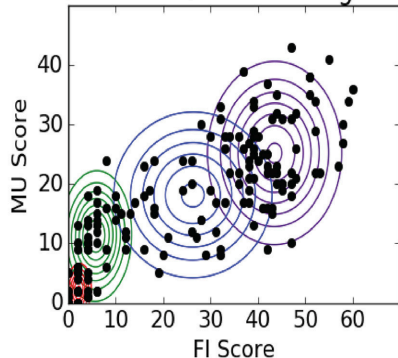
FIGURE 12 | EM contour plot.

participant's CPS sub-skills using evidence gathered from an online simulation/game. We showed how we can take the granular evidence gathered from the conversation flow and simulation/game activity data and map that onto our performance level estimates of latent variables, such as CPS skills. These higher level constructs are driven by CPS subject matter expert tagging and tunable conditional probability tables. This methodology creates a model that can be inspected at any time during the game to provide a probability-based estimate of participant ability. As we move forward with this work we can use

this model to start to build more sophisticated simulation/game interactions that could change adaptively, based on our real-time estimate of ability. For example, if we see participants are showing evidence of low feature identification we can add cues/tips to help them in this facet of interaction.

While the real-time Bayesian evidence tracing has proven useful in generating actionable insights for an individual participant during a game, our clustering work reported here has addressed our need to also compare across games. Our application of k-means gave us the ability to quickly characterize all games in the study and to group similar gameplays with each other, thus yielding different game profiles. Using K-NN we are able to treat these clusters as queryable sets that allow us to find participants that had similar evidence patterns of CPS sub-skills. In applying our Gaussian mixture model we were able to generate a more flexible cluster characterization of each game that can allow for partial cluster membership in more than 1 game profile.

We are working on the next iteration of our Circuit Runner game using the methods and results we have reported here. In our future work we are considering the integration of Bayesian evidence tracing with an application of adaptive conversation flows. We are also incorporating new instruments that will provide more demographics/data on the participants, such as a HEXACO assessment of personality and the results of a CPS questionnaire. We are also considering human-human CPS interaction scenarios that could feature scripted or open-ended conversations.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., and Wu, A. Y. (1998). An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *JACM* 45, 891–923. doi: 10.1145/293347.293348

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., et al. (2004). A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *J. Personal. Soc. Psychol.* 86:356. doi: 10.1037/0022-3514.86.2.356

Bauckhage, C., Drachen, A., and Sifa, R. (2015). Clustering game behavior data. *IEEE Trans. Comput. Intel. AI Games* 7, 266–278. doi: 10.1109/TCIAIG. 2014.2376982

Bayes, M., and Price, M. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philos. Trans.* 53, 370–418. doi: 10.1098/rstl.1763.0053

Bazaldua, D., Khan, S., von Davier, A. A., Hao, J., Liu, L., and Wang, Z. (2015). "On convergence of cognitive and non-cognitive behavior in collaborative activity," in *Proceedings of the 8th International Conference on Educational Data Mining* (Madrid), 496–499.

Bergner, Y., Andrews, J. J., Zhu, M., and Kitchen, C. (2015). "Agent-based modeling of collaborative problem solving," in *Paper presented at the 10th Annual INGRoup Conference* (Pittsburgh, PA).

Camara, W., O'Connor, R., Mattern, K., and Hanson, M. A. (2015). *Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success.* Act research report series. *ACT Inc.*

Canossa, A. (2013). "Meaning in gameplay: filtering variables, defining metrics, extracting features and creating models for gameplay analysis," in *Game Analytics*, eds M. Seif El-Nasr, A. Drachen, and A. Canossa (London: Springer), 255–283.

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User Adap. Inter.* 4, 253–278. doi: 10.1007/BF01099821

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., and Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment.* Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.

Desmarais, M. C., and Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User Adap. Inter.* 22, 9–38. doi: 10.1007/s11257-011-9106-8

Gow, J., Baumgarten, R., Cairns, P., Colton, S., and Miller, P. (2012). Unsupervised modeling of player style with lda. *IEEE Trans. Comput. Intel. AI Games* 4, 152–166. doi: 10.1109/TCIAIG.2012.2213600

Griffin, P., McGaw, B., and Care, E. (2012). *Assessment and Teaching of 21st Century Skills.* Dordrecht: Springer.

Hao, J., Liu, L., von Davier, A. A., Kyllonen, P., and Kitchen, C. (2016). "Collaborative problem solving skills versus collaboration outcomes: findings from statistical analysis and data mining," in *Proceedings of the 9th International Conference on Educational Data Mining* (Raleigh, NC).

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., and Borysewicz, P. B. (2017). "Collaborative problem solving measures in the programme for international student assessment (pisa)," in *Innovative Assessment of Collaboration*, eds A. A. von Davier, M. Zhu, and P. C. Kyllonen (Cham: Springer), 95–111.

Kerr, D. (2015). Using data mining results to improve educational video game design. *J. Educ. Data Mining* 7, 1–17. Available online at: http://jedm. educationaldatamining.org/index.php/JEDM/article/view/JEDM048

Kerr, D., and Chung, G. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *J. Educ. Data Mining* 4, 144–182. Available online at: http://jedm. educationaldatamining.org/index.php/JEDM/article/view/25

Kerr, D., Chung, G. K., and Iseli, M. R. (2011). *The Feasibility of Using Cluster Analysis to Examine Log Data from Educational Video games.* Cresst report 790. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Khan, S. (2015). *Multimodal Behavioral Analytics for Intelligent Training and Assessment Systems.* The rutgers university.

Khan, S., Cheng, H., and Kumar, R. (2013). "A hierarchical behavior analysis approach for automated trainee performance evaluation in training ranges," in *Proceedings of the 7th International Conference on Augmented Cognition*, eds D. D. Schmorrow, C. M. Fidopiastis (Berlin; Heidelberg: Springer), 60–69.

LaMar, M. M. (2014). *Models for Understanding Student Thinking Using Data from Complex Computerized Science Tasks.* PhD thesis, University of California, Berkeley.

Levy, R. (2014). *Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments.* Cresst report 837. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Li, H., Munoz-Avila, H., Ke, L., Symborski, C., and Alonso, R. (2013). "Discovery of player strategies in a serious game," in *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing* (Palm Springs, CA).

McLachlan, G. J., and Basford, K. E. (1988). *Mixture Models. Inference and Applications to Clustering.* New York: Dekker: Statistics: Textbooks and Monographs.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., et al. (2014). *Psychometric Considerations in Game-Based Assessment.* GlassLab Report. GlassLab Research, Institute of Play.

Mislevy, R. J., Steinberg, L. S., and Almond, R. G. (2003). Focus article: on the structure of educational assessments. *Measure. Interdiscipl. Res. Perspect.* 1, 3–62. doi: 10.1207/S15366359MEA0101_02

Mislevy, R. J., Steinberg, L. S., Almond, R. G., and Lukas, J. F. (2006). "Concepts, terminology, and basic models of evidence-centered design," in *Automated Scoring of Complex Tasks in Computer-Based Testing*, eds D. M. Williamson, R. J. Mislevy, and I. I. Bejar (Mahwah, NJ: Routledge), 15–47.

NCES (2017). *Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress.* NCES.

Oliveri, M. E., Lawless, R., and Molloy, H. (2017). A literature review on collaborative problem solving for college and workforce readiness. *ETS Res. Rep. Ser.* doi: 10.1002/ets2.12133. [Epub ahead of print].

Orkin, J., and Roy, D. (2011). "Semi-automated dialogue act classification for situated social agents in games," in *Agents for Games and Simulations II*, ed F. Dignum (Berlin: Springer), 148–162.

Paulhus, D. L., Vazire, S., Robins, R. W., Fraley, R., and Krueger, R. F. (2007). The self-report method. *Handb. Res. Methods Personal. Psychol.* 1, 224–239.

PISA, OECD (2013). *Results: Excellence Through Equity: Giving Every Student the Chance to Succeed.* (Vol. ii). PISA, OECD.

Robins, R. W., and John, O. P. (1997). The quest for self-insight: theory and research on accuracy and bias in self-perception. *Handb. Pers. Psychol.* 649–679. doi: 10.1016/B978-012134645-4/50026-3

Romero, C., González, P., Ventura, S., Del Jesús, M. J., and Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: a practical application using moodle data. *Exp. Syst. Appl.* 36, 1632–1644. doi: 10.1016/j.eswa.2007.11.026

Sánchez, J., and Olivares, R. (2011). Problem solving and collaboration using mobile serious games. *Comput. Educ.* 57, 1943–1952. doi: 10.1016/j.compedu.2011.04.012

Shute, V. J., Hansen, E. G., and Almond, R. G. (2008a). You can't fatten a hog by weighing it-or can you? Evaluating an assessment for learning system called aced. *Int. J. Artif. Intel. Educ.* 18, 289–316. Available online at: https://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai18-4-02

Shute, V. J., Ventura, M., Bauer, M., and Zapata-Rivera, D. (2008b). Monitoring and fostering learning through games and embedded assessments. *ETS Res. Rep. Ser.* 2008, i–32. doi: 10.1002/j.2333-8504.2008.tb02155.x

Smith, T. S. (2011). *Unsupervised Discovery of Human Behavior and Dialogue Patterns in Data from an Online Game.* PhD thesis, Massachusetts Institute of Technology.

Soller, A., and Stevens, R. (2007). "Applications of stochastic analyses for collaborative learning and cognitive assessment," in *Advances in Latent Variable Mixture Models*, eds G. R. Hancock and K. M. Samuelsen (Charlotte, NC: IAP), 217–253.

Steiner, I. (1972). *Group Processes and Productivity.* Cell Biology. New York, NY: Academic Press.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci.* 1:801.

Sung, H.-Y., and Hwang, G.-J. (2013). A collaborative game-based learning approach to improving students' learning performance in science courses. *Comput. Educ.* 63, 43–51. doi: 10.1016/j.compedu.2012.11.019

Teófilo, L. F., and Reis, L. P. (2013). Identifying player\'s strategies in no limit texas hold\'em poker through the analysis of individual moves. arXiv preprint arXiv:1301.5943.

Thurau, C., and Bauckhage, C. (2010). "Analyzing the evolution of social groups in World of Warcraft®," in *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on* (Copenhagen: IEEE), 170–177.

VanLehn, K. (2008). "Intelligent tutoring systems for continuous, embedded assessment," in *The Future of Assessment: Shaping Teaching and Learning*, ed C. A. Dwyer (New York: Routledge), 113–138. Available online at: http://www.public.asu.edu/~kvanlehn/Not%20Stringent/PDF/06Dwyer.pdf

von Davier, A. A. (2017). Collaborative educational assessments [special issue]. *J. Educ. Measure.* 54, 1–141. doi: 10.1111/jedm.12129

von Davier, A. A. (2015). "Virtual and collaborative assessments: Examples, implications, and challenges for educational measurement," in *Invited Talk Presented at the Workshop on Machine Learning for Education, International Conference of Machine Learning* (Lille).

von Davier, A. A., and Halpin, P. F. (2013). Collaborative problem solving and the assessment of cognitive skills: psychometric considerations. *ETS Res. Rep. Ser.* 2013, i–36. doi: 10.1002/j.2333-8504.2013.tb02348.x

von Davier, A. A., van der Schaar, M., and Baraniuk, R. (2016). "Workshop on machine learning for education," in *International Conference of Machine Learning* (New York, NY).

von Davier, A. A., Zhu, M., and Kyllonen, P. C. (2017). *Innovative Assessment of Collaboration.* Cham: Springer.

Zhang, M., Hao, J., Li, C., and P., D. (2015) "Classification of writing styles using keystroke logs: a hierarchical vectorization approach," in *Paper Presented at International Meeting of the Psychometric Society* (Beijing).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Determining Factors for Stress Perception Assessed with the Perceived Stress Scale (PSS-4) in Spanish and Other European Samples

Miguel A. Vallejo*, Laura Vallejo-Slocker, Enrique G. Fernández-Abascal and Guillermo Mañanes

Faculty of Psychology, Universidad Nacional de Educación a Distancia, Madrid, Spain

Objective: Stress perception depends on cultural and social aspects that vary from one country to another. One of the most widely disseminated methods of assessing psychological stress is the Perceived Stress Scale (PSS-4). Therefore, in order to identify these factors and their impact on mental health, the present study compares the PSS-4 results among three European countries (Great Britain, France and Spain). This study focuses on PSS-4 results within a Spanish sample to determine: (1) normative data, reliability and validity of PSS-4 in a Spanish sample and (2) how stress perception changes depending on cultural and social factors.

Methods: The data were obtained from a website representing a service of a smoking cessation program, the study represented a service that was open to all individuals. The number of participants were 37,451. They reported their age, gender, nationality, marital status, education and employment status, and completed two psychological questionnaires (PPS-4 and the anxiety and depression scales of the Symptom Checklist-90-Revised, SCL 90-R).

Results: The PSS-4 scores could differentiate between relevant sociodemographic variables (such as sex, age, nationality, marital status, education, parental status, employment status, and income class). The PSS-4 scores showed a positive correlation with the SCL 90-R anxiety and depression scales. The normed values for interpreting the PSS-4 scores are presented. The PSS-4 showed adequate internal consistency and reliability.

Conclusions: The PSS-4 is a useful instrument for assessing stress perception levels in the general population in different countries. Its internal consistency is sufficient for a 4-item scale.

Keywords: stress, perceived perception stress, normative values, psychometric, cross-cultural assessment

INTRODUCTION

Stress is an important reference point in health studies and it is related to both an individual's general health status and different illnesses, including mental disorders, cancer, cardiovascular disease, drug abuse, chronic diseases, etc. Since stress is a cross-cultural symptom for many different types of health problems, understanding stress across different sociodemographic, cultural and

social groups could help to prevent stress-related problems and major health concerns worldwide. The prevalence of mental disorders in Europe shows the importance of considering stress variables. Key reasons include mental disorders frequently being observed in females (Dedovic et al., 2009), the unemployed, persons who never married, youngers or countries where their living (ESEMeD, 2004). The cultural factors, including the social support, are fundamental to understanding how the people perceive stress and how they cope with it (Kim et al., 2008). These differences are present in European countries as UK, Germany, France, Netherlands, Spain and Belgium and show how the effect of the stress is different (Cao et al., 2016). Thus, evaluating the way in which individuals perceive stressful situations in their lives is critical for the quantification of psychological stress in health and disease worldwide. In particular, this research is necessary for understanding, preventing, and treating many health problems beyond national boundaries.

One of the most widely disseminated methods of assessing psychological stress has been the Perceived Stress Scale (PSS; Cohen et al., 1983). This self-report scale generates a global stress score that is based on general questions rather than focusing on specific experiences. This scale could be useful to compare stress perception in different countries. With this scale, subjects are asked to evaluate the previous month before the time of the self-report. The PSS was originally a 14-item scale, although the scale was reduced to 10 items (Cohen and Williamson, 1988) when 4 poorly performing items were identified and removed. Additionally, this was further reduced to 4 items for use in situations in which measurements must be obtained quickly. The PSS-4 scale has a clear advantage in terms of the time required to complete and ease of use, and this assessment is easy to complete on the Internet (Herrero and Meneses, 2006; Mañanes et al., 2016). The main reason for choosing the PSS-10 instead of the PSS-4 is not reliability (Cronbach α), as several studies have shown a reliability level of $\alpha = 0.67$ for the PSS-10 (Leung et al., 2010) versus $\alpha = 0.82$ for the PSS-4 (Mitchell et al., 2008). Although the PSS-4 shows better internal consistency than the PSS-10, the magnitude of the difference is more dependent on the study characteristics as opposed to the PSS scale (Lee, 2012).

These PSS scales have been studied, translated and adapted to several languages and countries, which on one hand demonstrates the relevance of this instrument and on the other hand, demonstrates the importance of stress among different societies. Warttig et al. (2013) published normative data for the PSS-4 in an English-speaking sample (population of several Primary Care Trusts from England and Wales), and Lesage et al. (2012) provided similar data from a French sample (workers selected from several occupational health care centers of the North of France). These authors published the norms in means and standard deviation (SD) for gender, group ages, ethnicity and other relevant variables. Additionally, these data allow for the comparison of different studies and samples. Lee

(2012) reviewed the psychometrical data in 19 studies that used the PSS-14, PSS-10, and PSS-4 scales. He concluded that the PSS scales show acceptable psychometric properties, albeit with differences related to population diversity, such as gender, age, parental status and other sociodemographic variables. Other studies (Cohen and Williamson, 1988; Herrero and Meneses, 2006; Karam et al., 2012; Lee et al., 2015; Ingram et al., 2016) used PSS-4 but did not offer norm values. Nevertheless, they informed of mean and standard deviation of a general sample, male, and female. It is useful to compare these data with the data obtained in normative studies.

Thus, it is necessary to evaluate these scales further. Particularly to obtain norms from broad representative samples and understand stress among cultures.

Within Europe, the PSS-4 has been tested among British and French samples. However, there are no studies of this scale in Spanish speaking countries. The PSS-4 may be useful for broadening knowledge about the determinants of stress because of its shortness, stress approach, Internet suitability and psychometric properties. Thus, the aim of this study was to examine a broad sample of the Spanish population to determine the following: (1) psychometric characteristics of the PSS-4 in terms of: normative data, reliability and validity in a Spanish sample and (2) how stress perception changes depending on cultural and social factors by comparing the results of this study with the British and French samples (a reference is made in relation to a Canadian and Korean study about stress assessed with PSS-4).

METHODS

The data for this study were obtained from the smoking cessation program website of the Universidad Nacional de Educación a Distancia (UNED) (https://www.apsiol.uned.es/dejardefumar) from October 2009 to November 2014. The study represented a service that was open to all individuals. The Bioethical Committee of the UNED approved the study. Before initiating the data introduction, informed consent was obtained from all participants.

A total of 37,451 people were interested in participating. Those who were interested in enrolling in the study were required to submit sociodemographic data, including age, gender, nationality, marital status, education and employment status, as well as two psychological questionnaires: the PPS-4 (Cohen et al., 1983; Cohen and Williamson, 1988) and the anxiety and depression scales of the Symptom Checklist-90-Revised (SCL 90-R) (Derogatis, 1977; González de Rivera et al., 1989). Additional details about the study have been published elsewhere (Mañanes and Vallejo, 2014; Mañanes et al., 2016). All the data were obtained from the Web application before the smoking cessation program began. There was not a selection of data. All the data registered was accepted.

The PSS-4, as well as, the SCL 90-R had already been tested in Spanish speaking samples in on-line interventions. We used the PSS-4 adapted by Herrero and Meneses (2006). The scale has four items in only one scale. These four items were: in the last month,

---

(1) *how often have you felt that you were unable to control the important things in your life?*

(2) *how often have you felt confident about your ability to handle your personal problems?*

(3) *how often have you felt that things were going your way?*

(4) *how often have you felt difficulties were piling up so high that you could not overcome them?*

A high score indicates a high perception of stress. The Cronbach's α coefficient = 0.72; and a Principal Component analyses show one factor that explains a 54% of the variance. Herrero and Meneses (2006) used a point scale ranging from 1 to 5 instead of the original scale used by Cohen (0–4). We used Cohen's original scale (0, never; 1, almost never; 2, sometimes; 3, fairly often; 4, very often) to score the 4-item scale. The SCL 90-R was also presented in Spanish on the Internet with a Cronbach's α coefficient = 0.97 (Vallejo et al., 2008).

The broad sample of people included in the study allowed us to examine the utility and psychometric structure of the PSS-4 in a Spanish population and to compare these data with other studies that have provided desegregated data from this questionnaire. An alpha level of 0.05 was used for the statistical test. An analysis of variance (ANOVA) was used to analyze the differences between the subgroups of the sample. A *t*-test was used to compare the results of this study with other studies when appropriate. The Pearson correlation coefficient was used to relate the quantitative variables.

To evaluate the PSS-4 psychometric properties, a reliability analysis was performed. We calculated the Cronbach's α coefficient, the Spearman Brown split-half reliability coefficient, and the standard error of measurement (SEM). A validity analysis included a correlation between the PSS-4 scores and the anxiety and depression SCL 90-R scores as well as the differences between the PSS-4 scores and the different relevant variables.

We performed a Principal Component analysis to determine the PSS-4 structure. We selected an Oblimin rotation as proposed by Lesage et al. (2012). Eigenvalues above 1 were retained.

RESULTS

Sample Description

The analysis of the sample ($n = 37,451$) showed a distribution similar to the Spanish population, as obtained through the Spain Census. **Table 1** shows the characteristics of the sample, which comprised 47.2% male and 52.1% female subjects. The mean age in the sample (38.9 years) was very close to that of the Spanish population (38.8 years). In the sample, 91.1% of the subjects were Spanish, similar to the frequency of 90.5% in the general population; 1.7% represented citizens of other EU countries, and 7.2% were from outside the EU, likely, for language reasons, from Spanish-speaking countries. The marital status of the sample showed that 43.8% were married and 35.8% were single. In terms of education, 47.3% completed university studies, 18.9% had professional training and 20.9% had high school education; the remaining 12.9% had primary-level education. In terms of employment, 75.9% were currently working and 24.1% were unemployed. This distribution fits perfectly with that reported in

TABLE 1 | Sample characteristics.

	Total	N	%	Spain census 2014 %
Sex	Male	17,673	47.2	49.1
	Female	19,778	52.1	50.9
Age	Years	Mean = 38.9	SD = 10.3	Mean = 38.8
Nationality	Spanish	34,116	91.1	90.5
	Other EU	629	1.7	4.2
	Non-EU	7,541	7.2	5.3
Marital status	Single	13,393	35.8	
	Married	16,403	43.8	
	Separated	3,425	9.2	
	Living with partner	3,870	10.3	
	Widowed	360	0.9	
Education	Primary	4,844	12.9	
	High school	7,818	20.9	
	Professional training	7,070	18.9	
	University	17,719	47.3	
Employment status	Working	28,512	75.9	75.7
	Unemployed	8,939	24.1	24.3
Income class	Upper	223	0.6	
	Upper-middle	4,343	11.6	
	Middle	23,330	62.3	
	Lower-middle	7,241	19.3	
	Bottom	2,314	6.2	

N = 37,451. EU, European Union.

the Spain Census. Finally, in terms of income, 62.3% belonged to the middle class, with a normal curve distribution showing slow predominant movement to the lower socioeconomic levels.

PSS-4 Scale Analysis

Table 2 shows the PSS-4 item scores with the mean and standard deviation. The mean score for the PSS-4 was 5.43 with a SD of 2.96. The kurtosis and skewness values indicated a normal distribution: below 0.05 in the global scale and below 0.1 for each item.

The internal consistency of the scale was 0.74 (Cronbach's α). The Spearman-Brown split-half reliability coefficient was 0.76, and SEM was 1.53. According to this SEM, the interval indicating the true score was 5.34 ± 1.53 (3.81–6.87).

The Kaiser-Meyer-Olkin (KMO = 0.794) index from the exploratory factor analysis confirmed that the sampling is adequate. The analysis obtained only one factor, explaining 56.4% of the total variance. The factor loadings were 0.78, 0.63, 0.79, and 0.81 for items 1–4, respectively. **Table 3** shows the factor loadings for principal component analysis.

The PSS-4 scores showed a positive correlation with the SCL 90-R anxiety and depression scales: $r = 0.51$ and $r = 0.69$,

TABLE 2 | Means, standard deviations, kurtosis and skewness for items and total scores of the PSS-4.

PSS-4	Mean	SD	Kurtosis	Skewness
Item 1	1.17	1.02	−0.71	0.65
Item 2	1.36	1.04	−0.45	0.64
Item 3	1.55	0.88	−0.23	0.23
Item 4	1.35	1.02	−0.94	0.56
Total score	5.43	2.96	−0.14	0.41

N = 37,451.

respectively. The correlations were significant ($p < 0.05$) and indicated concurrent validity with the PSS-4. A difference in stress perception was identified between working and unemployed subjects: 5.16 vs. 6.33; $F = 10871.31$ (1, 37449), with $p < 0.001$. Other differences between the groups with known distinct levels of stress, such as gender, age, education, parental status or income class, are shown in **Table 4**. The PSS-4 was able to differentiate between these groups.

Norm Values

Table 4 shows the disaggregated data of the PSS-4 scores for the different groups that were identified within the sample. The groups included sex (male, female), age intervals, nationality (Spanish, other European Union countries (EU), non-EU), marital status (single, separated, living with partner, widowed), education level (primary, high school, professional training, university), parental status (with children, without children), employment status (working, unemployed) and income class (upper, upper-middle, middle, lower-middle, lower). The number of subjects with the mean, standard deviation and quartile 1 and 3 values are shown for each group.

Table 5 shows other studies which used PSS-4 and do not offer norm values. Nevertheless, they informed of mean and standard deviation of a general sample, male, and female. It is useful to compare these results with those obtained in our study.

DISCUSSION

To our knowledge, this study used the largest population sample to study the PSS-4. Our results provide the opportunity to (1) analyze how stress is perceived throughout different cultures by comparing our results with other studies; (2) use this scale in a wide sample of Spanish (one of the most widely spoken languages) obtaining new norms for interpreting the scores; and (3) add information about the psychometric characteristics of this instrument.

Mean

We obtained a mean of 5.43, which was identical to the result from the study by Lesage et al. (2012), which used the French version of the PSS-4 (see **Table 4**). Several differences exist between other studies. In the British study (Warttig et al., 2013), the mean was 6.11. This result was statistically different (see **Table 4**) and may have been due to the comparison between Spanish and British participants. It actually seems like the higher

TABLE 3 | Factor loadings for principal component analysis.

PSS-4	Communality	Factor loading
ITEM 1		
Unable to control the important things in your life	0.60	0.78
ITEM 2		
Confident about your ability to handle your personal problems	0.50	0.63
ITEM 3		
Things were going your way	0.61	0.79
ITEM 4		
Difficulties were piling up so high that you could not overcome them	0.65	0.81

N = 37,451. Eigenvalue = 2.26. Percentage of variance explained = 56.4%.

differences are between British and other EU nationals (Spanish and French people have lower PSS-4 scores, 5.4, than British people, 6.1). Moreover, when people in our study with other EU nationalities were compared to the British sample, the difference was not significant. It is possible that, in relation to stress, people with other EU nationalities in Spain (6.05) share common characteristics with British (6.11) thus making them respond similarly to British and differently from Spanish (5.38) and French (5.40) people.

Other studies have also reported different means for stress perception. **Table 5** shows these data with a correction when the original 0–4 scale was not used. Ingram et al. (2016) obtained higher scores (7.05), which may have been due to a lower socioeconomic level of the participants. In our study (see **Table 4**), scores increased with both lower income classes (7.4 for the bottom income class) and lower education levels (6.02 for primary level education). Lee et al. (2015) also obtained a higher mean (6.27) than could be explained due to cultural factors (Korean sample). A low mean (2.88) was presented by Karam et al. (2012), but this represented a specific population of pregnant women. The other two studies included in **Table 4** reported values near the 5.4 mean of our study and the French study. In the study by Herrero and Meneses (2006), there was no significant statistical difference, and in the case of the original study by Cohen and Williamson (1988), a small difference was identified.

In summary, the mean of 5.4 obtained in our study may constitute a valid reference for perceived stress in Spain. When the mean is considered in the interval that is limited by the SEM (3.81–6.87), all the means can be included, with the exception of those with specific values (income class, education, etc.). **Table 4** includes two columns with information about quartiles 1 and 3 to help with the interpretation of the data in each subsample.

Age

Stress scores showed a tendency to decrease with age, from 6.25 in ages <18 years to 4.88 in ages >65 years. Similar findings were obtained by Cohen and Williamson (1988) and Warttig et al. (2013). This trend was not apparent in the French study (Lesage et al., 2012), although this difference may have been due to the

TABLE 4 | Norms for the stratified scores of the PSS-4.

		Present study						French study (Lesage et al., 2012)[¥]				British study (Warttig et al., 2013)			
		n	M	SD	Q1[§]	Q3[§]	F(fd)	n	M	SD	t(fd)	n	M	SD	t(fd)
Total		37,451	5.43	2.95	3	7		501	5.40	2.90	0.23 (35950)	1,484	6.11	3.14	10,37 (38933)***
Sex	Male	17,673	5.25	2.91	3	7	132,19 (1, 37449)***	249	5.20	2.50	0.31 (17920)	491	5.56	3.04	2.23 (18162)
	Female	19,778	5.60	2.99	3	8		252	5.60	3.20	0.00 (20028)	993	6.38	3.15	7.63 (20769)***
Age (years)	<18	229	6.25	2.95	4	8	24.16 (6, 37449)***	—	—	—	—	22	6.91	2.89	1.02 (249)
	18–29	7,180	5.73	2.98	4	8		125	4.70	2.00	5.65 (7303)***	409	6.66	3.23	5.68 (7587)***
	30–44	18,816	5.38	2.94	3	7		133	5.30	2.60	0.35 (18947)	596	6.05	3.16	5.11 (19410)***
	45–54	8,547	5.38	2.94	3	7		129	5.90	3.20	1.83 (8584)	264	5.69	3.04	1.63 (8809)
	55–64	2,368	5.08	2.93	3	7		114	5.60	2.40	2.23 (2480)	143	5.76	2.87	2.75 (2509)
	>65	311	4.88	2.90	3	7		—	—	—	—	50	5.32	2.86	1.01 (359)
Nationality	Spanish	34,116	5.38	2.94	3	7	52.16 (2, 37449)***	501	5.40	2.90	0.15 (34615)	1,484	6.11	3.14	8.79 (35598)***
	Other EU	629	6.05	2.95	4	8		501	5.40	2.90	3.71 (1128)***	1,484	6.11	3.14	0.42 (2111)
	Non-EU	7,541	5.9	3.08	4	8		501	5.40	2.90	3.72 (8040)***	1,484	6.11	3.14	2.36 (9023)
Marital status	Single	13,393	5.69	2.96	4	8	85.28 (4, 37449)***	162	5.50	3.20	0.75 (13553)				
	Married	16,403	5.13	2.89	3	7		337	5.30	2.70	1.14 (16738)				
	Separated	3,425	5.85	3.09	4	8		—	—	—	—				
	LWP	3,870	5.45	2.96	3	7		—	—	—	—				
	Widowed	360	5.51	2.99	3	7		—	—	—	—				
Education	Primary	4,844	6.20	3.12	4	8	1684.26 (3, 37449)***	176	5.80	2.90	1.79 (5018)				
	HS	7,818	6.33	2.99	4	8		149	5.10	3.10	2.07 (7965)				
	PT	7,070	5.52	2.86	3	7		53	5.50	2.30	0.06 (7121)				
	University	17,719	5.11	2.88	3	7		123	5.10	2.60	0.04 (17840)				
Parental status	C	19,758	5.32	2.96	3	8	63.01 (1, 37449)***	335	5.60	2.80	1.81 (20089)				
	NC	17,692	5.56	2.94	3	7		166	4.90	3.00	2.82 (17856)***				
Employment status	Working	28,512	5.16	2.85	3	7	1097.31 (1, 37449)***								
	Unemployed	8,939	6.33	3.12	4	8									
Income class	Upper	223	3.82	2.86	2	5	591.75 (4, 37449)***								
	Upper-middle	4,343	4.50	2.84	2	6									
	Middle	23,330	5.18	2.80	3	7									
	Lower-middle	7,241	6.24	2.97	4	8									
	Bottom	2,314	7.40	3.24	5	10									

***p > 0.001, **p > 0.01, *p > 0.05, Bonferroni adjusted for multiple comparisons, LWP, living with partner; PT, professional training; HS, high school; C, with children; NC, without children.

[¥] The age intervals and the education/occupational statuses were adapted to allow for comparisons.

[§] A score equal or lower than Q1 value indicates that this person belongs to the group of people with lower stress perception. He belongs to 25% of the people with lower stress perception. A score equal or higher than Q3 value indicates that this person belongs to 75% of the people with higher stress perception. He belongs to the group of people with higher stress perception.

TABLE 5 | PSS-4 scores in different studies.

		N	M	SD	Scale	Range	M (0–16)
Present study	Total	37,451	5.43	2.95	0–4	0–16	5.43
	Male	17,673	5.25	2.91	0–4	0–16	5.25
	Female	19,778	5.60	2.99	0–4	0–16	5.60
Ingram et al., 2016[a]	Total	225	11.05	3.41	1–5	4–20	7.05*
	Male	135	10.77	3.59	1–5	4–20	7.77*
	Female	90	11.44	3.23	1–5	4–20	8.44*
Karam et al., 2012[b]	Female	217	2.88	2.93	0–4	0–16	2.88*
Herrero and Meneses, 2006[c]	Internet	262	9.25	2.37	1–5	4–20	5.25
Lee et al., 2015[d]	Male	159	6.27	2.13	0–4	0–16	6.27*
	Female	243	7.22	2.41			7.22*
Cohen and Williamson, 1988[e]	Male	946	4.20	2.80	0–4	0–16	4.20*
	Female	1,784	4.70	3.10			4.70*

*p > 0.001; [a]USA, Low socioeconomic; [b]USA and Canada, Pregnant; [c]Spain; [d]Korea; [e]USA.

different intervals used to classify these variables. Several theories offer reasons for this decline of stress with age, from the selectivity of positive aspects to reduced physical reactivity due to physical and health limitations (Carstensen et al., 1999, 2006).

Nationality

Nationality is one way to consider cultural differences, which may be why the French and Spanish scores are more similar to one another than the British scores. In our study, the lower scores were provided by Spanish people (5.38), and these scores increased (5.9) when the subjects were Spanish-speaking (non-EU citizens) or had other EU nationalities (6.05) (likely immigrants). A similar increase was shown in the study by Warttig et al. (2013) related to ethnicity, including Caucasians compared to African-American, mixed and Asian subjects (with the exception of Chinese participants). The sensitivity of the PSS-4 to these cultural and ethnicity factors is a positive aspect of the scale and supports the results of other studies (Geronimus et al., 2006). Being able to assess stress patterns within a particular cultural environment is necessary to understand the nature of stress.

Marital Status

The lowest PSS-4 scores were obtained for married subjects (5.13) while the highest were obtained by separated subjects (5.85). The distribution obtained in our study showed that living with a partner had a positive effect on reducing stress. The same trend was found in the French study. Single and widowed individuals were in the middle of the score range. The protection/support hypothesis in favor of married individuals has been supported by various studies (Coombs and Fawzy, 1982). Additionally, population data show poor mental health for unmarried and widowed individuals compared to married individuals or those living with partners (Lindström and Rosvall, 2012).

Education

In terms of education, subjects with the lowest education identified the highest amount of stress (6.20) while those with

the highest about of education seem to identify the least (5.11). Lesage et al. (2012) obtained a similar result (5.80–5.10, respectively) while subjects with professional training or a High School education scored very similarly (5.52 and 5.62, respectively). Comparing the educational level of participants across different countries is a difficult task. The role of the educational level has not been acknowledged in the stress bibliography when considering the general population. Fiocco et al. (2007) found that people with lower levels of education were more stressed in reaction to the Trier Social Stress Test than individuals with higher levels of education. Instead of focusing on the importance of the education level, other variables, such as social or parental support, may mediate the education level influence on stress (Parkes et al., 2015).

Parental Status

Parental status is another familiar factor that may be related to stress. In our study, people with children showed low (5.32) PSS-4 scores compared to individuals without children (5.56). These data suggest that having children is a protective factor against stress. However, the French study (Lesage et al., 2012) showed a different result, in which subjects with children obtained a higher PSS-4 score (5.60) than those without children (4.90). In this case, having children was identified as a vulnerability factor for stress, and a similar result was obtained in Canada (Muhammad and Gagnon, 2009). The difference observed in our study may be explained in sociocultural terms, as it is possible that the social support of an extended family has a protective role in stress perception.

Employment Status

Unemployed subjects showed higher PSS-4 scores than employed individuals. This effect is congruent with many studies related to this issue. Economic and noneconomic factors are likely responsible for this difference (Frasquilho et al., 2016), and there is a clear independent effect of unemployment on mental distress (Backhans and Hemmingsson, 2012).

Income Class

The PSS-4 scores clearly reflected the six income classes defined in our study, from the lower class (7.40) to the upper class (3.82). These results confirmed previous studies showing a close relationship between stress and low income. Indeed, having a low income is a predictor of a variety of psychological problems (DeCarlo Santiago et al., 2011), and the perception of stress overlays other variables, such as gender or age (Panjwani et al., 2016).

The validity of the PSS-4 was confirmed in our study by the positive correlation with depression and anxiety scales of SCL 90-R, although the principal reason for recognizing the validity of this questionnaire is the power to differentiate between groups that are under different levels of stress. The internal consistency of the PSS-4 was similar to the values reported by other studies. Our alpha of 0.74 is within the range of those found by Cohen and Williamson (0.60, 1988) and Mitchell et al. (0.82, 2008). The Spearman Brown split-half reliability coefficient obtained was close to that obtained by Mitchell et al. (2008), and the SEM was slightly lower.

The factor structure obtained through the principal component analysis supports the one factor solution of the PSS-4. Similar results were obtained by others (Cohen and Williamson, 1988; Mitchell et al., 2008; Lesage et al., 2012), but others have questioned whether these four items really load onto two factors (Ingram et al., 2016).

In summary, our results show that the PSS-4 is useful for assessing stress in the general population in Spain and to establish comparisons with different countries. Our study, using a sample of 37,451 subjects, enabled the differentiation and facilitation of norms between subjects according to several sociodemographic variables. The results obtained are close to those obtained with other samples. Although cultural factors are important, the effect in different contexts exists when socioeconomic factors are considered. To this extent, not being able to compare all categories defined in our study to equivalent data in other samples was considered a limitation for the analysis of stress across cultures. Finally, the internal consistency of the PSS-4 is sufficient for a 4-item scale and for gathering information on the perception of stress.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study protocol was approved by the Bioethical Committee of the UNED. All participants gave written informed consent when they entered the study.

AUTHOR CONTRIBUTIONS

MV and GM designed the study and the aplication web needed to performing it. LV-S and EF-A constructed the conceptual framework of the work. MV and LV-S performing the data retrival and statistical analysis and preparing the initial draft of the manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

REFERENCES

Backhans, M. C., and Hemmingsson, T. (2012). Unemployment and mental health–who is (not) affected? *Eur. J. Public Health* 22, 429–433. doi: 10.1093/eurpub/ckr059

Cao, C., Zhu, C., and Meng, Q. (2016). An exploratory study of inter-relationships of acculturative stressors among Chinese students from six European union (EU) contries. *Int. J. Intercult. Rel.* 55, 8–19. doi: 10.1016/j.ijintrel.2016.08.003

Carstensen, L. L., Isaacowitz, D. M., and Charles, S. T. (1999). Taking time seriously. A theory of socioemotional selectivity. *Am. Psychol.* 54, 165–181. doi: 10.1037/0003-066X.54.3.165

Carstensen, L. L., Mikels, J. A., and Mather, M. (2006). "Aging and the intersection of cognition, motivation, and emotion," in *Handbook of the Psychology of Aging, 6th Edn.*, eds J. E. Birren, K. W. Schaire, J. E. Birren, and K. W. Schaire (Amsterdam: Elsevier), 343–362.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A global measure of perceived stress. *J. Health Soc. Behav.* 24, 385–396. doi: 10.2307/2136404

Cohen, S., and Williamson, G. (1988). "Perceived stress in a probability simple of the United States," in *The Social Psychology of Health,* eds S. Spacapan and S. Oskamp (Newbury Park, CA: Solid Action on Globalization and Environment), 31–68.

Coombs, R. H., and Fawzy, F. I. (1982). The effect of marital status on stress in medical school. *Am. J. Psychiat.* 139, 1490–1493. doi: 10.1176/ajp.139.11.1490

DeCarlo Santiago, C., Wadsworth, M. E., and Stump, J. (2011). Socioeconomic status, neighborhood disadvantage, and poverty-related stress: prospective effects on psychological syndromes among diverse low-income families. *J. Econ. Psychol.* 32, 218–230. doi: 10.1016/j.joep.2009.10.008

Dedovic, K., Wadiwalla, M., Engert, V., and Pruessner, J. C. (2009). The role of sex and gender socialization in stress reactivity. *Dev. Psychol.* 45, 45–55. doi: 10.1037/a0014433

Derogatis, L. R. (1977). *SCL 90-R Symptom Checklist 90 Revised.* Minneapolis, MN: NCS Pearson.

ESEMeD (2004). Prevalence of mental disorders in Europe: results ffrom the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatr. Scand.* 109 (Suppl. 420), 21–27. doi: 10.1111/j.1600-0047.2004.00327.x

Fiocco, A. J., Joober, R., and Lupien, S. J. (2007). Education modulates cortisol reactivity to the trier Social stress test in middle-aged adults. *Psychoneuroendocrinology* 32, 1158–1163. doi: 10.1016/j.psyneuen.2007.08.008

Frasquilho, D., de Matos, M. G., Marques, A., Gaspar, T., and Caldas-de-Almeida, J. M. (2016). Distress and unemployment: the related economic and noneconomic factors in a sample of unemployed adults. *Int. J. Public Health* 61, 1–8. doi: 10.1007/s00038-016-0806-z

Geronimus, A. T., Hicken, M., Keene, D., and Bound, J. (2006). "Weathering" and age patterns of allostatic load scores among blacks and whites in the United States. *Am. J. Public Health* 96, 826–833. doi: 10.2105/AJPH.2004.060749

González de Rivera, J. L., Derogatis, L. R., De Las Cuevas, C., Garcia Marco, R., Rodríguez Pulido, F., and Monterrey, A. L. (1989). *The Spanish Version of the SCL 90-R. Normative Data in the General Population.* Towson: Clinical Psychometric Research.

Herrero, J., and Meneses, J. (2006). Short Web-based versions of the perceived stress (PSS) and Center for Epidemiological studies-Depression (CESD) Scales: a comparison to pencil and paper responses among internet users. *Comput. Hum. Behav.* 22, 830–846. doi: 10.1016/j.chb.2004.03.007

Ingram, P. B., Clarke, E., and Lichtenberg, J. W. (2016). Confirmatory factor analysis of the perceived stress scale-4 in a community sample. *Stress Health* 32, 173–176. doi: 10.1002/smi.2592

Karam, F., Bérard, A., Sheehy, O., Huneau, M., Briggs, G., Chambers, C., et al. (2012). Reliability and validity of the 4-item Perceived Stress Scale among pregnant women: results from the OTIS antidepressants study. *Res. Nurs. Health* 35, 363–375. doi: 10.1002/nur.21482

Kim, H. S., Sherman, D. K., and Taylor, S. E. (2008). Culture and social support. *Am. Psychol.* 63, 518–526. doi: 10.1037/0003-066X

Lee, E. H. (2012). Review of the psychometric evidence of the Perceived Stress Scale. *Asian Nurs. Res.* 6, 121–127. doi: 10.1016/j.anr.2012.08.004

Lee, E. H., Chung, B. Y., Suh, C. H., and Jung, J. Y. (2015). Korean versions of the Perceived Stress Scale (PSS-14, 10 and 4): psychometric evaluation in patients with chronic disease. *Scand. J. Caring Sci.* 29, 183–192. doi: 10.1111/scs.12131

Lesage, F.-X., Berjot, S., and Deschamps, F. (2012). Psychometric properties of the French versions of the Perceived Stress Scale. *Int. J. Occup. Med. Environ. Health* 25, 178–184. doi: 10.2478/s13382-012-0024-8

Leung, D. Y., Lam, T., and Chan, S. S. (2010). Three versions of Perceived Stress Scale: validation in a sample of Chinese cardiac patients who smoke. *BMC Public Health* 10:513. doi: 10.1186/1471-2458-10-513

Lindström, M., and Rosvall, M. (2012). Marital status, social capital and health locus of control: a population-based study. *Public Health* 126, 790–795. doi: 10.1016/j.puhe.2012.06.001

Mañanes, G., and Vallejo, M. A. (2014). Usage and effectiveness of a fully automated, open-access, Spanish Web-based smoking cessation program: randomized controlled trial. *J. Med. Internet Res.* 16:e111. doi: 10.2196/jmir.3091

Mañanes, G., Vallejo, M. A., and Vallejo-Slocker, L. (2016). Demographic, psychological and smoking characteristics of users of an on-line smoking cessation programme in the Spanish language. *Gac Sanit.* 30, 18–23. doi: 10.1016/j.gaceta.2015.07.004

Mitchell, A. M., Crane, P. A., and Kim, Y. (2008). Perceived stress in survivors of suicide: psychometric properties of the Perceived Stress Scale. *Res. Nurs. Health* 31, 576–585. doi: 10.1002/nur.20284

Muhammad, A., and Gagnon, A. (2009). Why should men and women marry and have children? Parenthood, marital status and self-perceived stress among Canadians. *J. Health Psychol.* 15, 315–325. doi: 10.1177/1359105309353216

Panjwani, N., Chaplin, T. M., Sinha, R., and Mayes, L. C. (2016). Gender differences in emotion expression in low-income adolescents under stress. *J. Nonverb. Behav.* 40, 117–132. doi: 10.1007/s10919-015-0224-6

Parkes, A., Sweeting, H., and Wight, D. (2015). Parenting stress and parent support among mothers with high and low education. *J. Fam. Psychol.* 29, 907–918. doi: 10.1037/fam0000129

Vallejo, M. A., Mañanes, G., Comeche, M. I., and Díaz, M. I. (2008). Comparison between administration via internet and paper-and-pencil administration of two clinical instruments: SCL 90-R and GHQ-28. *J. Behav. Ther. Exp. Psychiat.* 39, 201–208. doi: 10.1016/j.jbtep.2007.04.001

Warttig, S. L., Forshaw, M. J., South, J., and White, A. K. (2013). New, normative, English-sample data for the Short Form Perceived Stress Scale (PSS-4). *J. Health Psychol.* 18, 1617–1628. doi: 10.1177/1359105313508346

TIMSS 2011 Student and Teacher Predictors for Mathematics Achievement Explored and Identified via Elastic Net

Jin Eun Yoo*

Department of Education, Korea National University of Education, Cheongju, South Korea

A substantial body of research has been conducted on variables relating to students' mathematics achievement with TIMSS. However, most studies have employed conventional statistical methods, and have focused on selected few indicators instead of utilizing hundreds of variables TIMSS provides. This study aimed to find a prediction model for students' mathematics achievement using as many TIMSS student and teacher variables as possible. Elastic net, the selected machine learning technique in this study, takes advantage of both LASSO and ridge in terms of variable selection and multicollinearity, respectively. A logistic regression model was also employed to predict TIMSS 2011 Korean 4th graders' mathematics achievement. Ten-fold cross-validation with mean squared error was employed to determine the elastic net regularization parameter. Among 162 TIMSS variables explored, 12 student and 5 teacher variables were selected in the elastic net model, and the prediction accuracy, sensitivity, and specificity were 76.06, 70.23, and 80.34%, respectively. This study showed that the elastic net method can be successfully applied to educational large-scale data by selecting a subset of variables with reasonable prediction accuracy and finding new variables to predict students' mathematics achievement. Newly found variables via machine learning can shed light on the existing theories from a totally different perspective, which in turn propagates creation of a new theory or complement of existing ones. This study also examined the current scale development convention from a machine learning perspective.

Keywords: machine learning, elastic net, regularization, penalized regression, TIMSS, mathematics achievement

INTRODUCTION

Apparently, the Lee Sedol vs. AlphaGo match last year shocked the world; the 18-time world champion was defeated by AI (Artificial Intelligence). Although AI had beaten human champions in chess and Jeopardy (a game show), in 1997 and 2011, respectively, the game of "Go" had been considered formidable for AI to conquer, partly due to the game's close to infinite number of cases. Simply speaking, the recent triumphs of AI such as AlphaGo are made possible primarily by machine learning techniques, which is training machines (computers) to learn through algorithms as well as data, and therefore to search for optimum solutions. Specifically, AlphaGo used policy and value networks of neural networks as well as Monte Carlo tree search algorithms to find the next best move within given time frame and ultimately to win the game (Silver et al., 2016).

However, the machine learning method, neural networks, is infamous for its overfitting problems, as parameters can be exponentially increasing with multiple layers. Regularization techniques control the growth of coefficients and thus are used to solve overfitting problems. Regularization can be carried out with penalized regression in statistics. Penalized regression techniques such as Ridge, LASSO (Least Absolute Shrinkage and Selection Operator), and elastic net have been widely applied to various fields of study including computer science/engineering (Keerthi and Shevade, 2007; Sun et al., 2012; Wang et al., 2013; Youyou et al., 2015; Zhou et al., 2015; Shen et al., 2016), biology/medicine (He and Lin, 2010; Nie et al., 2010; Yang et al., 2010b; Li et al., 2011; Waldmann et al., 2013; Wang et al., 2017a), and finance (Kim and Swanson, 2014; Wu et al., 2014; Borke, 2017; Wang et al., 2017b).

Penalized regression has been popular with big data analyses, especially in situations where there are many variables and few observations, so-called "large p, small n" problems (Schäfer and Strimmer, 2005; Zou and Hastie, 2005; Zhao et al., 2009; He and Lin, 2010; Waldmann et al., 2013). However, large-scale educational data such as TIMSS (Trends in International Mathematics and Science Study) can also benefit from penalized regression with its hundreds of variables and thousands of participants. Previous TIMSS studies employed few indicators in their models selected based on theories and literature review, although TIMSS provides hundreds of student and teacher variables which have been collected after multiple experts' evaluations also along with theories and literature review. This may be partly due to the fact that conventional statistical methods have difficulty handling hundreds of variables in one model, resulting in convergence and/or overfitting problems. Moreover, many TIMSS research focused on some student variables only, and a handful of studies dealt with student and teacher variables in one statistical model, although it is a well-accepted fact that teachers play a crucial role in students' performance.

Therefore, throwing the hundreds of TIMSS student and teacher variables in one model and selecting variables with machine learning techniques can shed light on the existing theories and literature, especially if not yet investigated variables are found to be important via this approach. For instance, this study newly found students' internet connection at home, car possession at home, and teacher specialization in language/reading as predictors for mathematics achievement. Teacher variables such as collaboration in planning and preparing instructional materials and parental support perceived by teachers were not also frequently investigated in previous research.

According to a systematic review on TIMSS studies by Drent et al. (2013), there had been scant grade four studies, and most studies were on western or top-performing countries. Particularly, Korea has been one of the top-performing countries, but was not one of the well-studied countries (Drent et al., 2013). As an attempt to fill the gap in research, this study explored student and teacher variables relating to students' mathematics achievement, using Korean 4th graders and their teachers as the sample.

To reiterate, the main goal of this study was to find a prediction model for students' mathematics achievement out of hundreds of TIMSS student and teacher variables. Elastic net, the selected machine learning technique in this study, handled the inevitable non-convergence and overfitting problems resulting from considering hundreds of variables in one statistical model as well as multicollinearity problems commonly encountered in social science data. Relatedly, applying machine learning techniques such as elastic net to educational large-scale data also has implications to the current scale development convention, as items do not need to be parceled primarily for scale development and data reduction purposes.

In the next section, penalized regression methods including elastic net are explained in more detail along with bias-variance tradeoff, other variable selection methods, and cross-validation. Before directly moving to elastic net, ridge, and LASSO are introduced as its predecessors, and their limitations are discussed as well. In a nutshell, elastic net is preferable to ridge and LASSO for its variable selection feature and its strength in multicollinearity problems, respectively, and therefore was the selected machine learning technique of this study to analyze the TIMSS data.

REGULARIZATION

Bias-Variance Trade-Off

In regression, the primary goal is to find the coefficient estimate close to the parameter. MSE (mean squared error) is used to evaluate this goal. Another goal which has not yet received deserved attention particularly in social sciences would be to find a model which fits future observations well. PE (Prediction Error) serves for evaluating this second goal, and is a well-documented topic in statistics (Hastie et al., 2009). PE comprises MSE and irreducible error. As the "irreducible" error is literally irreducible, attention is paid to variance and squared bias, the two components of MSE.

Notably, bias and variance trade-off each other. When a model becomes more complex, it picks up local structures of the data, and the bias of the coefficients gets lower, but the variance gets higher. As a result, overfitting may occur. In contrast, a simpler model increases the bias, but decreases the variance. Conventional ordinary least squares (OLS) or maximum likelihood (ML) methods have focused on obtaining unbiased estimates and lowering the variance among the unbiased estimates. On the other hand, regularization techniques focus on decreasing the overall MSE, by finding biased but lower-variance estimates.

Variable Selection

Variable selection is an important issue in data analysis, especially when there are many variables. Variable selection methods such as best-subset selection, forward selection, and backward elimination have been used for model construction, but they have weaknesses, compared to penalized regression methods (Hastie et al., 2009). To be more specific, best-subset can be applied to data with no more than 30–40 variables, and backward elimination has difficulty in dealing with the "large p, small n"

problems. Both forward selection and backward elimination as discrete methods lack stability in model construction, as they either include or remove a variable one by one.

On the other hand, penalized regression methods, also called as shrinkage methods, continuously penalize coefficients with a regularization parameter. This continuous nature of penalized regression methods are known to produce more stable models than the aforementioned discrete methods. The most widely used penalized regression methods are ridge, LASSO, and elastic net.

Ridge

Ridge was originally invented for multicollinearity problems (Hoerl and Kennard, 1970), but now is also well-known as an early penalized regression method. Suppose response variable y is estimated with X matrix of N observations and P predictors, and the X^TX matrix is singular. Ridge puts an additional lambda value in the diagonals of the X^TX matrix, and therefore the previously singular X^TX matrix becomes invertible. Using the ridge regression, coefficients get shrunken due to the additional lambda.

$$\hat{\beta} = argmin \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \right\}$$

(1)

The penalty parameter, λ, determines the amount of regularization. It is easily shown that the λ value of 0 turns the equation into least squares estimation. In Equation (1), the first term in the right-hand side is ordinary least squares part, and the second term is the penalty function, which is called as L_2 penalty. It is notable that ridge does not perform variable selection.

LASSO

LASSO (Least Absolute Shrinkage and Selection Operator), invented by Tibshirani (1996), was designed to obtain a model of higher prediction accuracy and better interpretation than models from ordinary least squares methods. Compared to ridge, LASSO equation has a different penalty function, which is L_1 penalty (Equation 2). While ridge sums up squared coefficients, LASSO uses the sum of absolute values. In a two-dimensional coefficients space, ridge's penalty constraint is shaped as a disk, and that of LASSO a diamond (Hastie et al., 2009). The error contours of L_1 LASSO penalty has corners, and if the elliptical contours hit the corner, the corresponding coefficient becomes zero (Hastie et al., 2009). Therefore, let alone better interpretability, LASSO has the advantage of variable selection over ridge, which has huge practical implications especially in the "large p, small n" problems, frequently occurring in big data analyses.

$$\hat{\beta} = argmin \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\}$$

(2)

As was with ridge, the penalty parameter, λ, controls the amount of regularization. Larger values of λ shrink the coefficients more,

and smaller values of λ makes the equation closer to least squares estimation. Unlike ridge, the estimation of LASSO does not provide closed forms, and therefore quadratic programming is required (Tibshirani, 1996).

Elastic Net

While LASSO is capable of variable selection, ridge is famous for its performance with multicollinearity (Zou and Hastie, 2005). Bridging between LASSO and ridge, elastic net exerts the advantages of the two, utilizing the L_1 and L_2 penalties of LASSO and ridge, respectively, in one equation. That is to say, elastic net not only selects variables, but also performs better than LASSO with collinear data (T. Hastie, personal communication, February, 9, 2017; R. Tibshirani, personal communication, February 1, 2017). As social science data such as TIMSS cannot be free from multicollinearity problems, particularly with its hundreds of variables, elastic net was the chosen method of this paper. The objective function of elastic net for Gaussian family is presented in Equation (3).

$$\hat{\beta} = argmin \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 \right.$$
$$\left. + \lambda \sum_{j=1}^{P} (\alpha |\beta_j| + (1-\alpha)\beta_j^2) \right\}$$

(3)

More specifically, elastic net has two parameters, λ and α. The regularization (or penalty) parameter, λ, functions the same as with ridge and LASSO. The new tuning parameter, α, bridges between ridge and lasso. If α is 1, the equation equals to LASSO (equation 2), and α of 0 returns the ridge equation (Equation 1). Therefore, the value of α in-between 0 and 1 determines whether the model is closer to ridge or LASSO, taking advantage of both ridge and LASSO.

Cross-Validation (CV)

The penalty parameter, λ, determines the amount of regularization and thus model complexity. Choosing the right value of λ which minimizes prediction error is an essential part in penalized regression. K-fold cross-validation (CV) is a common approach to obtain the penalty parameter. K-fold CV partitions training data into K sets of equal size. K is typically chosen to be 5 or 10. For each fold ($k = 1, 2, \ldots K$), the model is fitted with the training set excluding the k-th fold, and the fitted values are obtained for the k-th fold. This is repeated for every k-th fold, and each fold's CV error is calculated. The average of the K folds' CV errors serves as the overall CV error (Equation 4), and its standard error is also obtained.

$$CV(\hat{f}, \lambda) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i, \lambda))$$

(4)

The best model which minimizes prediction error can be specified with CV, but the "one-standard-error rule" is typically employed to get the most parsimonious model (Hastie et al., 2009). This means that the least complex model is chosen among the models within one-standard error range of the best model. Aforementioned, a good model not only should fit the data well, but also fit new data well. In machine learning, a good model

is obtained with "training" data, and evaluated with "test" (or an independent set of) data. Consequently, the corresponding λ is applied to the test data, and the model is evaluated with prediction accuracy, specificity, and sensitivity of the test data.

METHODS

Data Characteristics

In estimating students' academic achievement, TIMSS employs multiple imputation due to its matrix-sampling booklet design. Multiple imputation as a Bayesian approach draws multiple imputed (or plausible) values. TIMSS provides 5 PVs (plausible values) for mathematics and science as well as their subdomains, respectively. PVs are continuous, and TIMSS also provides categorical benchmarks for all respective PVs in 5 levels: 1 (Below Low; PVs < 400), 2 (Low; 400 <= PVs < 475), 3 (Intermediate; 475 <= PVs < 550), 4 (High; 550 <= PVs < 625), and 5 (Advanced; PVs >= 625).

Grade 4 student and teacher datasets of TIMSS 2011 Korea were merged, using the syntax codes of IEA's IDB Analyzer (Version 3.2.17). Although the original student dataset consisted of 4,334 students, the merged dataset had a total of 4,771 observations. The difference of 437 was due to the fact that two teachers of the 437 students responded to the teacher questionnaire. This study kept the first observation of the duplicates, which resulted in the original number of 4,334 observations with a total of 586 variables.

Response Variable

After merging, each student's mathematics class was created using all the five mathematics categorical benchmarks (ASMIBM01 to ASMIBM05) via majority vote, an ensemble method by Breiman (1996). For instance, if a student's benchmark variables were 1, 1, 2, 2, 1, then the student's class was coded as 1. **Table 1** presents the majority vote results. There were 34 ties out of 4,334 Korean 4th graders. These ties were deleted from the subsequent analyses, and therefore the final sample size was 4,300.

As about half of the students were in the "Advanced" level (5), the first four levels were collapsed, generating a new benchmark variable, ASMIBM. ASMIBM was coded with the criterion whether students reached "Advanced" (coded as "1"; 42%) or not (coded as "0"; 58%), and was used as the response variable of this study.

Explanatory Variables

Starting with the dataset of 586 variables on 4,300 4th graders and their teachers, irrelevant, or duplicate variables were removed and missing data was handled. Firstly, irrelevant variables relating to IDs (e.g., school ID, student ID, etc.), file maintenance (e.g., date of testing, file creation date, etc.), and weights (e.g., total student weight, etc.) were deleted from the explanatory variable pool. For duplicate variables such as students' gender (ITSEX, ASBG01) and birth information (e.g., ITBIRTHM, ASBG02A), variables of students' responses (e.g., ASBG01, ASBG02A) were removed.

Secondly, values such as omitted or invalid, logically not applicable, or not administered were marked as missing. The "omitted" responses came from the respondents' carelessness or unwillingness to answer the question. The "not applicable/administered" responses mainly resulted from the fact that TIMSS participating countries had great diversity in their educational systems. The missing rate of each variable was calculated, and 201 variables with their missing rates over 10% were removed from the dataset. This was a necessary step to maintain at least half of the original samples after listwise deletion.

Notably, except the newly created benchmark variable, ASMIBM, which served as the response variable of this study, all the other benchmark variables and PVs were excluded from the elastic net model. Inclusion of these academic performance variables would have dominated the model, which conveys little useful information to predict students' math achievement.

Among the explanatory variables, binary variables were dummy-coded so that girls and "Yes" responses (e.g., home possessions, etc.) were coded as 1; boys and "No" were coded as 0. Likert-like response variables such as number of books at home and computer use frequency were treated as continuous and coded as the original values.

Cross-Validation

After listwise deletion, this data cleaning process resulted in 2,353 4th graders (55% of the original data) with 163 student and teacher variables. For model validation, the observations were randomly split into training and test data sets with the conventional ratio of 7:3. The training data was used for model construction, and test data was for model evaluation and generalization.

Particularly, the response variable, ASMIBM, was used as the stratifying variable to keep the rate of "Advanced" vs. "Others" in the training and test datasets. **Table 2** presents numbers of students in each level for the training and test data sets.

Ten-fold CV (cross-validation) was used in this study. The training data were randomly split into 10-folds, and the model was fitted and evaluated using a range of λ values. The "one-standard-error rule" was employed for the most parsimonious model (Hastie et al., 2009), and the corresponding λ was identified and applied to the test data. Finally, the most

TABLE 1 | Majority vote result with 2011 TIMSS Korean 4th graders' math.

Level	1	2	3	4	5	Ties	Total
Observations	9	99	684	1,771	1,737	34	4,334

TABLE 2 | Training and test data.

	1 (Advanced; 42%)	0 (Others; 58%)
Data (n = 2,353)	995	1,358
Training data (n = 1,647)	696	951
Test data (n = 706)	299	407

parsimonious model was evaluated with prediction accuracy, specificity, and sensitivity of the test data.

	Predicted as positive (Advanced)	Predicted as negative (Others)
Actual +	TP (True Positive)	FN (False Negative)
Actual −	FP (False Positive)	TN (True Negative)

The prediction accuracy was calculated as the sum of true positives and true negatives divided by the total. For instance, prediction accuracy of 70% indicates that the model correctly predicts 70 out of 100 new students' status (Advanced or Others). Sensitivity indicates the probability that a data point actually true is classified as true, and was calculated as true positives divided by the sum of true positives and false negatives ($= \frac{TP}{TP+FN}$). Specificity indicates the probability that a data point actually false is classified as false, and was calculated as true negatives divided by the sum of false positives and true negatives ($= \frac{TN}{FP+TN}$).

Elastic Net With Logistic Regression

As the response variable was dichotomous (G = 1; Advanced), a logistic regression model was utilized as the analysis model (Equation 5). As standardization of variables is necessary in penalized regression (Hastie et al., 2009), coefficients were estimated using Equation (6) after standardization.

$$\log \frac{P(G=1 \mid X=x)}{P(G=0 \mid X=x)} = \beta_0 + \beta^T x \qquad (5)$$

$$\max \left\{ \beta_0 \epsilon R, \beta \epsilon R^p \right\} \left[\frac{1}{N} \sum_{i=1}^{N} \log P(G_i|x_i) - \lambda \sum_{j=1}^{P} \left(\alpha \left| \beta_j \right| \right. \right.$$
$$\left. \left. + (1-\alpha) \beta_j^2/2 \right) \right] \qquad (6)$$

All the programs were written in R 3.1.1. Specifically, the "glmnet" library was used (Friedman et al., 2017). The elastic net tuning parameter, α, was chosen as 0.5, as this value is known to perform well with correlated predictors (Hastie and Qian, 2016). To determine the penalty parameter, λ, a 10-fold CV was executed with cv.glmnet. The cv.glmnet package provides five types of measures for logistic regression models: misclassification error, AUC (Area Under the receiver operating characteristic Curve), binomial deviance, MSE (mean squared error), and MAE (mean absolute error).

Misclassification error is the proportion of misclassified cases among the total. AUC is the area under the ROC (Receiver Operating Characteristic) curve, and AUC of 1 indicates the perfect fit. Binomial deviance (or deviance) is a twice negative binomial log likelihood of the fitted model evaluated on the test data, and considered as an extension of the ordinary least squares' residual sum of squares in generalized linear models. MSE is the average of squared differences between actual and predicted values. MAE is the average of absolute differences between actual and predicted values. Compared to MAE, MSE penalizes large deviations more. AUC is compared to the baseline value of 0.5, and higher values of AUC indicate better model. All

the other measures are interpreted the opposite; lower values of misclassification error, deviance, MSE, and MAE indicate better performance.

The steps for the coefficient estimation were as the following. First, this study used all the five measures. The λ value of each measure was determined using the "one-standard-error rule" (Hastie et al., 2009). Second, five models with the corresponding λ values from the first step were obtained. Their accuracy, sensitivity, and specificity results with the test data were compared, and the measure of the best prediction accuracy was selected. Lastly, elastic net coefficients were obtained, using the λ value in the previous step.

RESULTS

Figure 1 shows coefficients' paths with increasing values of λ, the regularization tuning parameter. Each curve corresponds to a predictor. The numbers above the box indicate the numbers of non-zero coefficients with the corresponding $\log(\lambda)$ values on the X-axis. The Y-axis indicates the coefficients of predictors. We can see that the coefficients get close to 0 with increasing λ.

Figure 2 shows the 10-fold CV results of the five types of measures: misclassification error, AUC (Area Under Curve), deviance, MSE (Mean Squared Error), and MAE (Mean Absolute Error). As was with **Figure 1**, the numbers above the box indicate the numbers of nonzero coefficients with their corresponding $\log(\lambda)$ values on the X-axis. The Y-axis is in the unit of each measure. With all the measures except AUC, a lower value on the Y-axis indicates better performance. The vertical dotted lines in each plot are the upper and lower bounds of the one-standard-error rule. For instance, the fourth plot in **Figure 2** shows the MSE result, and the number of nonzero coefficients with the upper bound [larger $\log(\lambda)$] corresponds to 17. Therefore, a total of 17 variables were selected for the most parsimonious model with the one-standard-error rule using MSE. Among the five measures, MSE and misclassification error yielded the more parsimonious models, and MAE resulted in the least parsimonious model.

Table 3 presents prediction accuracy, sensitivity, and specificity results using the regularization parameter (λ) values of the 10-fold CV based on each measure. Each measure's number of variables and $\log(\lambda)$ were shown in **Figure 2**. Among the five measures, MSE showed the best performance, considering accuracy, sensitivity, and specificity of the test data as well as model parsimony. The regularization parameter, λ, was 0.0648 with the one-standard-error rule. This means that the MSE of the selected λ was within one standard error of the minimum value for the most regularized model. As results, 12 student and 5 teacher variables were identified out of the 162 predictors. The prediction accuracy, sensitivity, and specificity of the elastic net model with the test data were 76.06, 70.23, and 80.34%, respectively (**Table 3**).

Table 4 listed coefficients of selected variables from the elastic net model. Notably, penalized regression focuses on decreasing overall MSE by sacrificing the 'unbiasedness' of estimates. As the estimates of penalized regression are biased, the variance no
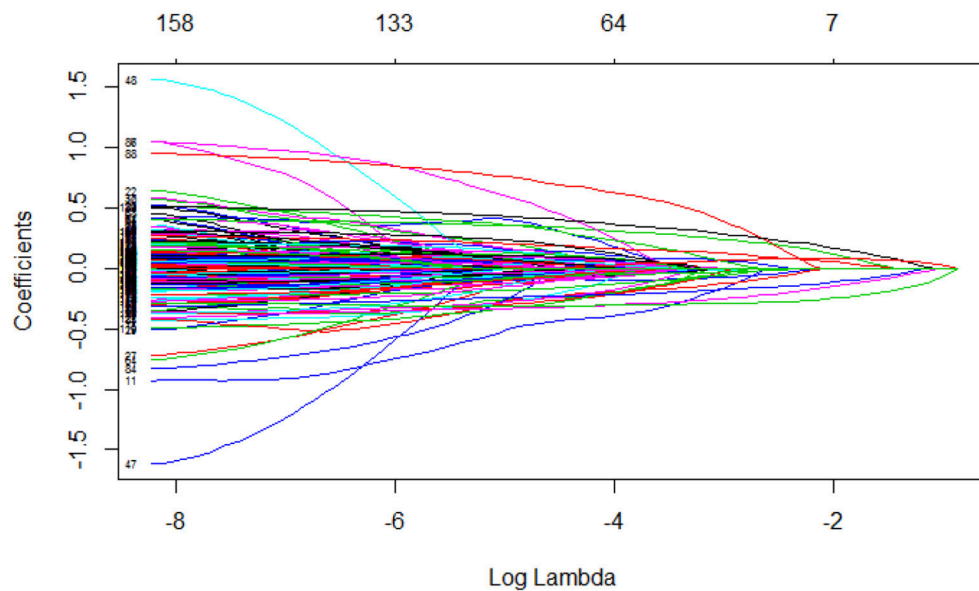
FIGURE 1 | Regularization parameter (λ) and corresponding coefficients.
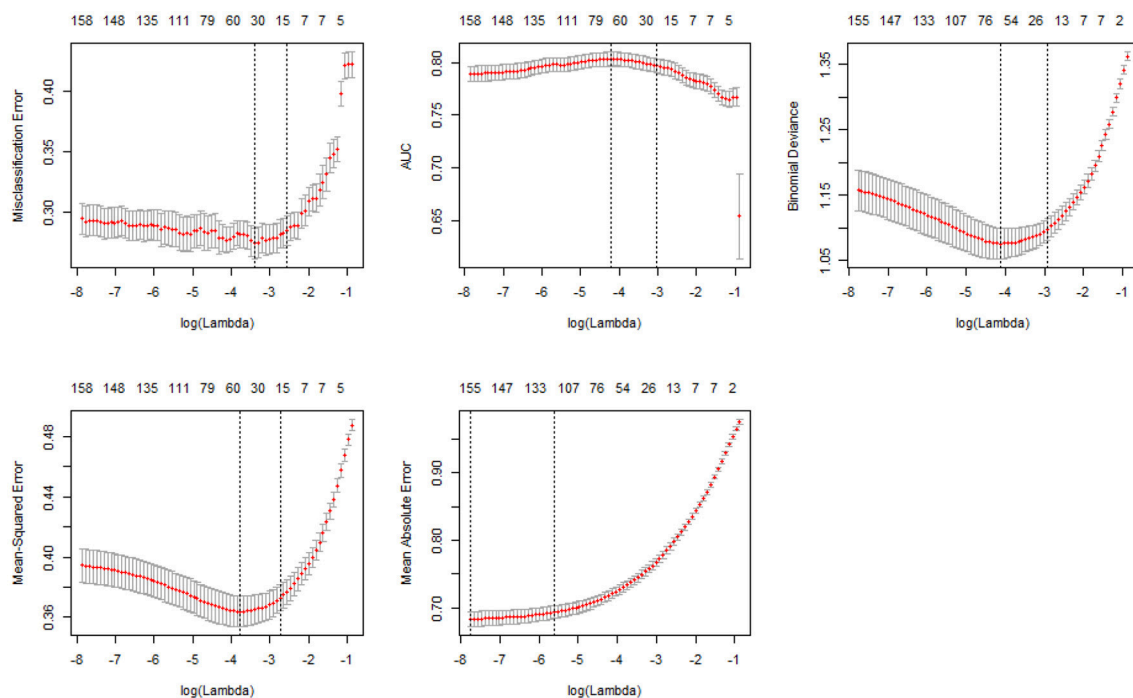
FIGURE 2 | Ten-fold CV results of five measure types.

longer equals to MSE. Although standard errors can be obtained via bootstrapping, standard errors only partially contribute to the MSE of penalized regression due to the biased estimates. As precise estimation of bias is nearly impossible with penalized regression (Goeman et al., 2016), precise estimation of MSE is also impossible. Therefore, it is typical that standard errors of coefficients are not provided with penalized regression.

As expected, students' math self-confidence was a crucial factor to their math achievement (**Table 4**). Out of the 12 student variables, 7 variables related to students' math self-confidence.

Specifically, the first five items measuring students' math self-confidence (ASBM03A to ASBM03E) were selected such as "I usually do well in mathematics," "Mathematics is harder for me than for many of my classmates," "I am just not good at mathematics," "I learn things quickly in mathematics," and "I am good at working out difficult mathematics problems." The other two items in the same scale (ASBM03F and ASBM03G), "My teacher tells me that I am good at mathematics," and "Mathematics is harder for me than any other subject" were not included in the model. The corresponding self-confidence composite score (ASBGSCM) and composite index (ASDGSCM) variables were also included in the elastic net model. One item on students' engagement in math lessons was selected: "My teacher

is easy to understand" (ASBM02C). The more positively students responded to these items, the more likely they achieved the Advanced level.

The remaining four student variables were about home resources including book possession (ASBG04), internet connection (ASBG05E), car possession (ASBG05F), and computer usage (ASBG06A) at home. The more books the students had at home or the more often they used computer at home, the more likely they achieved the Advanced level. Interestingly, students who had a car at home or internet connection at home performed better than those who did not. Particularly, the internet connection and students' math self-confidence index variables had the highest coefficients among the 17 selected variables to predict students' math achievement, followed by individual self-confidence items, and amount of books at home.

Aforementioned, five teacher variables were selected. Specialization in language/reading (ATBG05BC) was the only variable in the teacher background section ("About You"). Students whose teacher specialized in language/reading had lower chance of achieving the Advanced level, but math or science specialization was not included in the model. Three variables were selected out of the school characteristics ("About Your School"). High levels of parental support for student

TABLE 3 | Prediction accuracy, sensitivity, and specificity of five measures.

	Accuracy	Sensitivity	Specificity
Misclassification	73.23	56.67	84.39
AUC	73.83	59.07	83.77
Deviance	73.75	60.00	83.02
MSE	76.06	70.23	80.34
MAE	73.75	63.15	80.90

TABLE 4 | Selected variables, their labels, scales, and coefficients.

	Variable	Label	Scale	Coefficient
	Intercept			−1.262
1	ATBG05BC	Specialization/language-reading	1 (yes), 0 (no)	−0.060
2	ATBG06E	Parental support	1 (very high) to 5 (very low)	−0.060
3	ATBG06F	Parental involvement	1 (very high) to 5 (very low)	−0.005
4	ATBG07A	Sch/safe neighborhood	1 (agree a lot) to 4 (disagree a lot)	−0.098
5	ATBG10B	Interactions t/collaborate	1 (never or almost never) to 4 (daily or almost daily)	0.042
6	ASBG04	Amount of books in your home	1(0–10), 2(11–25), 3(26–100), 4(101–200), 5(200+)	0.214
7	ASBG05E	Internet connection	1 (yes), 0 (no)	0.287
8	ASBG05F	Car possession	1 (yes), 0 (no)	0.003
9	ASBG06A	How often use computer home	1 (everyday) to 4 (never)	−0.004
10	ASBM02C	Teacher is easy to understand	1 (agree a lot) to 4 (disagree a lot)	−0.030
11	ASBM03A	Usually do well in math	1 (agree a lot) to 4 (disagree a lot)	−0.230
12	ASBM03B	Harder for me than for others	1 (agree a lot) to 4 (disagree a lot)	0.255
13	ASBM03C	Just not good in math	1 (agree a lot) to 4 (disagree a lot)	0.103
14	ASBM03D	Learn quickly in mathematics	1 (agree a lot) to 4 (disagree a lot)	−0.024
15	ASBM03E	Good at working out problems	1 (agree a lot) to 4 (disagree a lot)	−0.163
16	ASBGSCM	Math self-confidence (score)		0.070
17	ASDGSCM	Math self-confidence (index)	1 (confident) to 3 (not confident)	−0.287

Teacher variables were followed by student variables, and variables were presented in the order of TIMSS questionnaires.

achievement (ATBG06E) or parental involvement in school activities (ATBG06F) perceived by teachers positively related to students' achievement. The more the teachers agreed with the statement that their school was located in a safe neighborhood (ATBG07A), their students had a higher chance of achieving the Advanced level. Particularly, the safe neighborhood variable had the highest coefficient among teacher variables, followed by parental support. Lastly, one teacher variable on teacher interaction was selected. The more the teacher collaborated with other teachers in planning and preparing instructional materials (ATBG10B), the higher the students' performance was.

DISCUSSION

Conventional statistical techniques such as Hierarchical Linear Model (HLM) and Structural Equation Model (SEM) value theories. After theory-laden variables are identified, researchers design sampling strategies, collect data, run analyses, and interpret variables of statistical significance. This has been the typical process of quantitative research, which was valid and acceptable when study design and data collection cost considerable time and/or money. However, with the advent of so-called big data era, researchers have access to enormous amounts of data, without individually spending time and money for data collection. The primary questions left relate to how to analyze the data. Machine learning can be one of the solutions. Particularly, machine learning methodology has been gaining increasing interest in the eLearning community, as conventional statistical methodology carries clear limits to LMS (Learning Management Systems) types of data analysis (Guha, 2017; Pappas, 2017).

This foremost study showed the possibility of extending the new methodology to educational large-scale data, TIMSS 2011. Specifically, elastic net, a regularization method, was employed among the machine learning techniques. Social science research does not need to be confined to conventional methodology. Employing machine learning techniques provides methodological and practical advantages over conventional statistical methods including new variable exploration and identification without convergence problems. Implications of this study are discussed, followed by conclusion.

Full Use of Data
First, this study made a full use of the wide range of TIMSS student and teacher data after data cleaning. This is virtually impossible with conventional methods such as HLM and SEM, as they fail to converge due to increasing number of parameters to be estimated. Partly related to this problem, previous research has been confined within a small set of variables or factors, depending on existing theories, statistical results, or a mix of the two. With machine learning and big (or large-scale) data, we may escape the researcher treadmill, and be able to find new important variables which have been ignored in the literature.

For instance, this study found internet connection and car possession at home were positively related to Korean 4th graders' math achievement. Particularly, car possession at home was a country-specific variable, and there was no TIMSS research using that variable as a predictor in math achievement. Internet

connection, another home possession variable, was also rarely investigated as an individual predictor. Car possession and internet connection at home along with book possession and computer use at home are indicators to household income. This result corroborates the fact that students' household income or socio-economic status closely relate to their math achievement.

While frequently studied TIMSS student variables such as math self-confidence and amount of books at home were also identified in the elastic net model, often studied TIMSS teacher variables relating to differentiation/ adaptive instruction, curriculum quality, or class climate were not selected in the model. Likewise, all the five teacher variables of this study were not listed as significant factors for the 4th graders in the Drent et al. (2013)'s systematic review. The levels of safety at school and parental involvement perceived by teachers were included in the Drent et al.'s 8th grader results, though.

Among the five selected teacher variables, teacher collaboration in planning and preparing instructional materials appears to be relatively instantly implemented in practice, as this is something teachers can do on their end. In contrast, the other teacher variables such as parental support, parental involvement, and school neighborhood safety all perceived by teachers as well as specialization in major are not variables that can be readily changed. Therefore, studies on teacher collaboration in instructional materials should be furthered in order to improve students' math achievement.

Although this new approach can provide researchers with methodological breakthrough and novel insights, its inherently data-driven approach may seem inappropriate from the conventional view. There are ways to incorporate researchers' prior knowledge in regularization techniques such as elastic net. For instance, the R glmnet library has a penalty factor function (p.fac). One can easily set the penalty as "0" for variables of theoretic importance which should be included in the model.

Scale Development
Second and related to the first point, scale development may not be necessary with regularization methods such as elastic net. Aforementioned, increasing number of parameters to be estimated can result in convergence problems in conventional methods, and therefore data reduction has been a major issue in statistical research. Although item parceling has been popular for this matter in psychological studies, "to parcel or not to parcel" also has been under debate (Little et al., 2002, 2013; Yang et al., 2010a; Marsh et al., 2013).

In fact, scale development by item parceling can be troublesome. Item parceling prevents convergence problems by summing or averaging a set of items and making an index (or composite) variable. However, this practice of summing or averaging assumes that the set of items are from a unidimensional trait and are equally contributing to measurement of that trait. Thus, high reliability of the scale is a necessity, but it is not always the case in practice. For instance, TIMSS 2011 reported Cronbach's alphas of several scales such as "Students engaged in mathematics lessons" (Martin and Mullis, 2012). Unfortunately, around 30% of the participating countries had the Cronbach's alphas below 0.50, and the majority had the alphas around 0.60.

The lowest alpha was that of Georgia, which was merely 0.21. This clearly violates the unidimensionality assumption required for scale development.

Another problem of item parceling relates to its differing item composition under the same label. That is to say, under the same labeling of a latent variable, different item parcels are often used depending on research. For instance, the latent variable, SES (Social Economic Status), is considered to be one of the most influential predictors to students' math achievement, and thus has been frequently studied in previous TIMSS research. However, different studies have used different combinations of items such as home resource items and parents' educational levels, although all the studies claimed that they indirectly measured "SES." Therefore, we have to be cautious in interpretation when item parcels are used.

One easy answer to these problems of item parceling is not-to-parcel. The convergence problem resulting from not-to-parcel can be solved by regularization methods such as elastic net. To reiterate, elastic net requires neither item-parceling for data reduction purposes, nor assumptions such as unidimensionality for item-parceling, but selects important variables out of hundreds of predictor candidates without convergence problems. Moreover, individual variable's effect is also identifiable with the selected variables' coefficients. That is to say, if an item is selected in the model with the highest coefficient among a set of items in a scale, then we can say that that item exerts the highest influence on prediction accuracy in the scale. This is not easily done with item parceling, as item parceling lumps individual items together.

Lastly, TIMSS item parceling is vulnerable to issues of missing data. TIMSS estimates scale scores for composite variables or indices such as students' self-confidence in math (ASBGSCM or ASDGSCM), if students responded to only two or more items in a given scale (P. Foy, personal communication, July 17, 2015). If there are missing responses, summing only available responses can result in different subsets of items and thus biased estimates (Schafer and Graham, 2002). To reiterate, different sets of variables used for scale score estimation, depending on the missingness patterns, may result in biased estimates.

Under this circumstance, listwise deletion is considered the better method than the available-data method (Allison, 2001). However, large-scale datasets such as TIMSS severely suffers from listwise deletion. Particularly, non-administered or non-applicable (NA) responses plague a number of TIMSS variables, especially from the teacher questionnaire. With listwise deletion, handling these NA responses as missing results in dramatic reduction in sample size and thus reduction in efficiency. This study removed variables of more than 10% missing, which retained about 55% of the original samples. If the missing rates were increased to 20 or 30%, only 15 and 7% of the original samples remained, respectively. Removing variables of more than 10% missing partly alleviated the sample size issues, but potentially important variables, especially in the teacher questionnaire, might have been removed as a result. Research should be furthered on missingness patterns, scale score estimation, and regularization methods such as elastic net, particularly in the context of educational large-scale data.

CONCLUSION

Research in the field of education has not yet paid enough, due attention to the recent big data/ machine learning techniques. Particularly, this study was one of the first studies to analyze educational large-scale data via elastic net. There can be disagreements on what constitutes educational big data, but educational large-scale data tentatively can serve the purposes of machine learning with its hundreds of variables and thousands of participants.

This study aimed to explore and identify possible sets of predictors using elastic net, currently one of the most popular machine learning techniques. A logistic regression model was employed to predict TIMSS 2011 Korean 4th graders' math achievement. Among 162 TIMSS variables, 12 student and 5 teacher variables were selected in the elastic net model, and its prediction accuracy was 76.06%. This means that the elastic net model of only 17 variables successfully predicted new students' mathematics class with 76.06% accuracy. Furthermore, this study was able to identify new predictors not yet investigated in previous research with conventional statistical methods. This study intentionally analyzed Korean 4th graders to fill the gap in the current TIMSS literature. Further machine learning studies with other TIMSS samples will help accumulate knowledge on students' math achievement, and consequently contribute to increasing students' math achievement.

It should be noted that penalized regression techniques focus on model prediction, not statistical significance. Therefore, predictors selected in penalized regression might not be statistically significant (T. Hastie, personal communication, January 28, 2017; R. Tibshirani, personal communication, January 28, 2017). Yoo and Rho (2017) employed group LASSO on social science large-scale data, and 15 out of 338 predictors were selected in the model. For comparison purposes, they constructed another model consisting of 83 predictors, based on literature review. Surprisingly, their model of only 15 predictors defeated the model of 83 predictors almost by 10%P in terms of prediction accuracy, specificity, and sensitivity. However, not all the 15 selected predictors in the group LASSO model were statistically significant.

In conclusion, it appears necessary to explore new variables to predict students' academic achievement and to examine the scale development convention via a machine learning technique. Newly found variables via machine learning can shed light on the existing theories from a totally different perspective, which in turn propagates creation of a new theory or complement of existing ones.

AUTHOR CONTRIBUTIONS

JY designed the study, cleaned the data, performed data analyses, and wrote the manuscript.

ACKNOWLEDGMENTS

REFERENCES

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Borke, L. (2017). *RiskAnalytics: An R Package for Real Time Processing of Nasdaq and Yahoo Finance Data and Parallelized Quantile Lasso Regression Methods (No. 2017-006)*. SFB 649 Discussion Paper.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 26, 123–140. doi: 10.1007/BF00058655

Drent, M., Meelissen, M. R. M., and van der Kleij, F. M. (2013). The contribution of TIMSS to the link between school and classroom factors and student achievement. *J. Curric. Stud.* 45, 198–224. doi: 10.1080/00220272.2012.727872

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2017). *Package 'glmnet.'* Available online at: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Goeman, J., Meijer, R., and Chaturvedi, N. (2016). *L1 and L2 Penalized Regression Methods.* Available online at: https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

Guha, S. (2017). *The Importance of Data Analytics in eLearning.* Available online at: https://elearningindustry.com/data-analytics-in-elearning-importance

Hastie, T., and Qian, J. (2016). *Glmnet Vignette.* Available online at: https://web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edn.* New York, NY: Springer.

He, Q., and Lin, D. Y. (2010). A variable selection method for genome-wide association studies. *Bioinformatics* 27, 1–8. doi: 10.1093/bioinformatics/btq600

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Keerthi, S. S., and Shevade, S. (2007). A fast tracking algorithm for generalized lars/lasso. *IEEE Trans. Neural Netw.* 18, 1826–1830. doi: 10.1109/TNN.2007.900229

Kim, H. H., and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence. *J. Econ.* 178, 352–367. doi: 10.1016/j.jeconom.2013.08.033

Li, W., Feng, J., and Jiang, T. (2011). IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* 18, 1693–1707. doi: 10.1089/cmb.2011.0171

Little, T. D., Cunningham, W. A., Shahar, G., and Widaman, K. F. (2002). To parcel or not to parcel: exploring the questions, weighing the merits. *Struct. Equat. Model.* 9, 151–173. doi: 10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, G., and Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychol. Methods* 18, 285–300. doi: 10.1037/a0033266

Marsh, H. W., Ludtke, O., Nagengast, B., and Morin, A. J. S. (2013). Why item parcels are (almost) never appropriate: two wrongs do not make a right-Camouflaging misspecification with item parcels in CFA models. *Psychol. Methods* 18, 257–284. doi: 10.1037/a0032773

Martin, M. O., and Mullis, I. V. S. (eds.). (2012). *Methods and Procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Available online at: http://timssandpirls.bc.edu/methods/t-context-q-scales.html#

Nie, F., Huang, H., Cai, X., and Ding, C. H. (2010). "Efficient and robust feature selection via join L2, 1-norms minimization," in *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 1813–1821. Available online at: http://www.citeulike.org/user/zzb3886/article/9315911

Pappas, C. (2017). *Machine Learning and Artificial Intelligence: The Future of eLearning.* Available online at: https://elearningindustry.com/machine-learning-artificial-intelligence-future-elearning

Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4, 1–30. doi: 10.2202/1544-6115.1175

Schafer, J. L., and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychol. Methods* 7, 147–177. doi: 10.1037/1082-989X.7.2.147

Shen, B., Liu, B. D., and Wang, Q. (2016). Elastic net regularized dictionary learning for image classification. *Multimed. Tools Appl.* 75, 8861–8874. doi: 10.1007/s11042-014-2257-y

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Sun, S., Huang, R., and Gao, Y. (2012). Network-scale traffic modeling and forecasting with graphical lasso and neural networks. *J. Transp. Eng.* 138, 1358–1367. doi: 10.1061/(ASCE)TE.1943-5436.0000435

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* 4:270. doi: 10.3389/fgene.2013.00270

Wang, J., Zhou, J., Wonka, P., and Ye, J. (2013). "Lasso screening rules via dual polytope projection," in *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems* (Lake Tahoe, NV). Available online at: https://dl.acm.org/citation.cfm?id=2999731#source

Wang, X., Nan, B., Zhu, J., Koeppe, R., and Frey, K. (2017a). Classification of ADNI PET images via regularized 3D functional data analysis. *Biostat. Epidemiol.* 1, 3–19. doi: 10.1080/24709360.2017.1280213

Wang, X., Zhong, J. Y., and Li, Z. Y. (2017b). Construction of quantitative transaction strategy based on LASSO and neural network. *Appl. Econ. Finan.* 4, 134–144. doi: 10.11114/aef.v4i4.2370

Wu, L., Yang, Y., and Liu, H. (2014). Nonnegative-lasso and application in index tracking. *Comput. Stat. Data Anal.* 70, 116–126. doi: 10.1016/j.csda.2013.08.012

Yang, C., Nay, S., and Hoyle, R. H. (2010a). Three approaches to using lengthy ordinal scales in structural equation models: parceling, latent scoring, and shortening scales. *Appl. Psychol. Meas.* 34, 122–142. doi: 10.1177/0146621609338592

Yang, C., Wan, X., Yang, Q., Xue, H., and Yu, W. (2010b). Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC Bioinformatics* 11:S18. doi: 10.1186/1471-2105-11-S1-S18

Yoo, J. E., and Rho, M. (2017). Predictor exploration via group LASSO: Focusing on middle school students' life satisfaction (written in Korean). *Stud. Korean Youth* 28, 127–149. doi: 10.14816/sky.2017.28.1.127

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1036–1040. doi: 10.1073/pnas.1418680112

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.* 37, 3468–3497. doi: 10.1214/07-AOS584

Zhou, Q., Chen, W., Song, S., Gardner, J. R., Weinberger, K. Q., and Chen, Y. (2015). "A reduction of the elastic net to support vector machines with an application to GPU computing," in *Proceedings of AAAI Conference on Artificial Intelligence* (Austin, TX), 3210–3216. Available online at: https://dl.acm.org/citation.cfm?id=2888163

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

A Non-linear Predictive Model of Borderline Personality Disorder Based on Multilayer Perceptron

Nelson M. Maldonato[1]*, Raffaele Sperandeo[2], Enrico Moretto[2] and Silvia Dell'Orco[2]

[1] Department of Neuroscience, Reproductive and Odontostomatological Sciences, University of Naples Federico II, Naples Italy, [2] Scuola in Psicoterapia Gestaltica Integrata, Torre Annunziata, Italy

Borderline Personality Disorder is a serious mental disease, classified in Cluster B of DSM IV-TR personality disorders. People with this syndrome presents an anamnesis of traumatic experiences and shows dissociative symptoms. Since not all subjects who have been victims of trauma develop a Borderline Personality Disorder, the emergence of this serious disease seems to have the fragility of character as a predisposing condition. Infect, numerous studies show that subjects positive for diagnosis of Borderline Personality Disorder had scores extremely high or extremely low to some temperamental dimensions (harm Avoidance and reward dependence) and character dimensions (cooperativeness and self directedness). In a sample of 602 subjects, who have had consecutive access to an Outpatient Mental Health Service, it was evaluated the presence of Borderline Personality Disorder using the semi-structured interview for the DSM IV-TR personality disorders. In this population we assessed the presence of dissociative symptoms with the Dissociative Experiences Scale and the personality traits with the Temperament and Character Inventory developed by Cloninger. To assess the weight and the predictive value of these psychopathological dimensions in relation to the Borderline Personality Disorder diagnosis, a neural network statistical model called "multilayer perceptron," was implemented. This model was developed with a dichotomous dependent variable, consisting in the presence or absence of the diagnosis of borderline personality disorder and with five covariates. The first one is the taxonomic subscale of dissociative experience scale, the others are temperamental and characterical traits: Novelty-Seeking, Harm-Avoidance, Self-Directedness and Cooperativeness. The statistical model, that results satisfactory, showed a significance capacity (89%) to predict the presence of borderline personality disorder. Furthermore, the dissociative symptoms seem to have a greater influence than the character traits in the borderline personality disorder e disease. In conclusion, the results seem to indicate that to borderline personality disorder development, contribute both psychic factors, such as temperament and character traits, and environmental factors, such as traumatic events capable of producing dissociative symptoms. These factors interact in a nonlinear way in producing maladaptive behaviors typical of this disorder.

Keywords: borderline personality, neural network, character and temperament variables, dissociative phenomena, personality disorders

INTRODUCTION

Borderline Personality Disorder (BPD) is one of the major challenges for contemporary psychopathology as far as understanding its pathogenesis, clinical manifestations and treatment methods are considered. In terms of etiopathogenesis, substantial epidemiological (Ball and Links, 2009) evidence shows that subjects with BPD have traumatic events in their preceding anamnestic history. According to recent work (Meares, 2014), it would seem that BPD has many similarities with dissociative disorders regarding the difficulty that patients with both types of disorders have in regulating vegetative phenomena and emotional activation status. Other relevant research (Cloninger, 2000) from the point of view of the pathogenetic investigation revealed a relationship between BPD and extreme temperament and character trait expressions. For example, research (Pukrop, 2002; Joyce et al., 2010) conducted using the Cloninger Temperament and Character Test (TCI) showed that low levels of Self-Directedness and Cooperativeness and high levels of Harm Avoidance and Novelty Seeking are related to the presence of personality disorders described in Cluster B of the DSM-IV-TR (American Psychiatric Association, 1996).

Dissociative phenomena (Sperandeo et al., 2018), on the other hand, commonly conceptualized as an interruption in the usually integrated functions of consciousness (Maldonato, 2009, 2014; Maldonato et al., 2011), memory, identity and perception of the environment (DSM-IV), have been widely associated, in the same way as BPD (Korzekwa et al., 2009; Meares, 2014), traumatic events, Post-Traumatic Stress Disorder (PTSD) (Yager, 1976; Spiegel et al., 1988; Carlier et al., 1996), sexual abuse during childhood and deficit in parental care (Brodsky et al., 1995; Zlotnick et al., 1996).

It is useful to clarify that when it comes to "dissociation" there is not necessarily a reference to a clinically defined syndrome; in fact, this term in its current use, in scientific literature, refers to both clinically relevant symptomatic phenomena and adaptive mental processes (Spiegel et al., 2011; Cantone et al., 2012).

In addition, according to Grabe, dissociation is a phenomenon which is closely related to the personality dimension. Indeed, in a study based on the Cloninger's Seven Factor model (Svrakic et al., 1993), it was discovered that: in women and men, the temperament and character inventory dimensions of Self-Transcendence, Self-Directedness, were predictors for the Dissociative Experience Scale scores. No temperament dimensions showed any significant predictive power. Such evidence, according to the authors, would confine the genesis of dissociative phenomena to the context of events caused by environmental stresses (Grabe et al., 1999).

From the analysis of current literature, an expression ultimately emerges that expresses the close relationship between traumatic events, dissociative phenomena and the dimensions of temperament and character in the genesis of BPD. However, these relationships are clearly non-linear so that none of the phenomena described can be considered as the first or the essential etiologic mechanism. They seem rather tied to each other in a complex network characterized by overlapping areas and recursive mechanisms.

The currently acclaimed model of the Borderline Personality Disorder (BPD) pathogenesis shows that Traumatic environmental events, interacting with the structural fragility of a subject's personality trigger dysfunctional processes in the area of cognition, emotions, interpersonal relationships, and social behavior.

The interaction between the personality structure and the traumatic environmental events is a very articulate and complex process and, at the moment, only superficially touched by scientific enquiry. Studies have only succeeded in establishing correlations between traits of fragile personality, traumatic events and BPD; in some cases, it has also been possible to describe, through regression methods, the relationship between the phenomena discussed in this study, but the effect size of these relationships is always rather limited. Although these explanatory models are of great importance for the evolution of knowledge, their prediction capacity is still functionally inadequate.

Moreover, the question of the quality of the interaction between psychological and environmental processes is completely unrealistic at this stage; specifically, it is important to question how it proceeds and develops, what are the factors that enhance or inhibit it, and how and to what extent it is possible to predict the emergence of BPD following the collapse of individual resilience systems. This last question is extremely relevant also for the possibility of activating effective primary and secondary preventions interventions.

In this regard, it is useful to refer to the theory proposed by Brosboom as a methodology for exploring the non-linear interactive processes typical of psychological disorders. Network theory, a recent epistemological model in psychopathology (Borsboom and Cramer, 2013), allows to conceptualize mental disorders as the result of direct interactions between symptoms. The biological, psychological and social mechanisms to which the individual is exposed in his life experience, according to this model, activate the emergence of a specific symptom and this triggers a process of amplifying the network of related psychopathological phenomena. For example, an angiogenic environmental stimulus may induce an alteration in sleep-wake rhythm that, if persistent, triggers asthenia and demotivation, followed by other typical symptoms of depressive disease (Kendler, 2012). If the relationship between the symptoms is strong enough, they can generate a level of feedback such as self-feeding the system. Although this pattern of psychopathology is not universally accepted, it is an effective attempt to explain the non-linear relationships between the psychological phenomena underlying the pathogenetic mechanism of a disorder (Borsboom, 2017).

Aims

The clinical relevance of BPD requires the search for effective prevention methods and hence the need for studies to gain not only certain evidence (as the current literature seems to outline) but also that it be materially usable in the development of instruments capable of predicting the onset of this disease with a clinically useful sensitivity and specificity.

The present study aims to investigate the quality of the pathogenetic process by improving the ability to predict, in a

single subject, the BPD on the basis of the evaluation instruments applicable in ordinary clinical contexts. For this purpose, it is necessary to use statistical analysis instruments suitable for modeling non-linear phenomena that favor the ability to predict the emergence of BPD by the interaction between environment and psychological phenomena.

INSTRUMENTS

Structured Clinical Interview for DSM (SCID-II)

SCID-II is an instrument used to diagnose personality disorders, both categorically (current or absent) or dimensionally. It consists of 119 items, having dichotomous answer modes Yes/ No. Every personality disorder is identified by a certain number of items that match DSM criteria for that specific diagnostic category. On the clinical interview model, the starting part consists of a brief overview that identifies the subject's normal behavior and relationships and allows to verify its introspection capabilities. A "3" score on a SCID-II item, provided by the clinician during the investigation, indicates that there is sufficient evidence that the feature described in the article is "pathological," "persistent," and "diffuse." "Pathological" indicates that the characteristic is outside the range of normal variation; "Persistent" refers to both frequency and duration (a score of "3" means that the feature has been present frequently during the last 5 years); "Diffused" indicates the presence of the feature in various contexts, such as at home and at work, or, in the case of items relating to interpersonal relationships, it comes out in different relationships. The interview was conducted by 5 trained researchers who achieved a high inter-rater reliability ($k = 0.81$) (First et al., 1997; Maffei et al., 1997).

Dissociative Experience Scale (DES)

DES is a self-report instrument for rapid compilation and processing that can assess the presence, quantity, and type of dissociative experiences without entering into the merit of diagnosis. It consists of 28 items arranged on an analog scale and scores vary from 0 to 100, for each item, and for the total score obtained from the average of the scores. The cut-off value indicating the presence of pathological dissociation concerns scores ≥ 20: scores greater than 20 are generally associated with a DSM-IV-TR dissociative disorder diagnosis, lower scores are commonly observed in both healthy subjects and in psychiatric patients in general.

The factorial analysis provided an understanding of the DES structure: for the Italian version, the three-factor model is the most frequently used one. According to this model, DES is composed of the following sub-scale:

- Dissociative Amnesia, which concerns actions of which the subject has no memory (item: 3, 4, 5, 8, 11);
- Imaginative Absorption And Involvement, indicating total immersion into an activity so as to become completely unaware of the surrounding environment (items 2, 14, 15, 17, 18, 20, 24);

- Depersonalization-Derealization, or altered perceptions of the self and the environment, such as feeling disconnected from one's own body, from their own thoughts, from their feelings (items 7, 12, 13, 21, 22, 23, 27, 28).

Subsequently, Waller et al. have selected 8 items of DES (item: 3, 5, 7, 8, 12, 13, 22, 27) by building a new subcategory that can identify the tendency to pathological dissociation, DEStax (Waller et al., 1996).

DES is a valid and reliable instrument for measuring dissociative experiences both in clinical and control samples and reveals a similar factorial structure in groups of psychiatric patients and normal subjects. Scores at DES-TOT over 20 are considered indicative of a pathological condition, but have no diagnostic value (Bernstein and Putnam, 1986; Fabbri Bombi et al., 1996).

Temperament and Character Inventory (TCI)

In order to come up with a diagnosis according to the Cloninger model, the instrument that the same author has developed can be used: Temperament and Character Inventory (TCI). The TCI is a self-report (self-report) questionnaire that, in its most complete version, is composed of 240 items in response to a dichotomy (true/false). Of these, 116 explore the 4 temperamental dimensions (NS, HA, RD, and P), 119 evaluate the 3 character dimensions (SD, C, and ST) and 5 are indicators of the presence of Personality Disorders. The sum of the items marked as "true" provides the raw scores of the seven scales. The raw scores are transformed into T standardized scores that, shown on a graph, provide a personality profile of the subject. This instrument includes questions about tastes, interests, emotions, responses, goals, and values. TCI results can be evaluated as raw score, T score, and scoring percentage, and a conversion table is provided between these three measures based on the score obtained in a standardization on a sample of 300 adults called a community sample; Cloninger says that this is representative of the general population and supports the reliability and structure of the TCI size. Temperament is considered to be the emotional heart of the personality. It includes four largely independent dimensions: (1) Novelty Seeking (NS), which represents behavioral activation in response to novelty and reward or punishment relief; (2) Harm Avoidance (HA), referring to behavioral inhibition in response to punishment or non-reward signals; (3) Reward Dependence (RD), reflecting socially rewarded behaviors; (4) Persistance (P) which describes persevering behavior despite fatigue and frustration. Character, however, is defined in terms of individual differences in the concept of self-experience that evolves throughout life in response to socio-cultural influences. It includes three dimensions: (1) Self Directedness (SD), which is the ability to adjust and adapt behavior to the needs of a situation in order to achieve personally chosen goals; (2) Cooperativeness (C), which expresses self-identification as an integral part of a more or less inclusive social group and the level of liking in relationships with others; (4) Self Transcendence (ST), which is associated with the ability to recall the past and to imagine the

future as the evolution of life's history, as well as experiencing unity with nature and developing spiritual values (Cloninger et al., 1993).

SAMPLE

The sample consists of 602 subjects of which 160 are diagnosed with Borderline Personality Disorder (BPD). The subjects belonging to the sample had an average age of 34.7 years of age with a standard deviation of 11.48; 266 were males and 336 were females; 42 subjects had completed elementary school, 119 had completed middle school, 326 had achieved high school diploma, 114 were graduates and only one participant lacked any qualification. Regarding the civil status, 327 persons are singles, 226 are married, 34 are separated or divorced, 15 are widowed. Finally, 419 subjects are employed in work or study, 174 are unemployed and 9 are retired.

METHODS

To achieve the aim of the study a feed-forward neural network type, that evolves through a process called Error Back Propagation (EBP), was used. This type of neural network called Perceptron Multilayer function as a powerful interpolation mathematical "trainer" system capable of calculating non-linear functions of any complexity from known input and output values used as examples for learning (Rumelhart et al., 1985).

Figure 1 depicts the Perceptron Multilayer Network EBP used in the study. This consisting of: an input layer of 5 input variables, of a hidden layer consisting of 10 nodes of computation and an output layer of consisting of two output nodes.

Two types of assessment were used as input variables; on the one hand, measurements (obtained with TCI) of the intensity of expression of the temperament and character traits that in the pertaining literature were systematically correlated with BPD, on the other the measurement (obtained with the DES) of the intensity of the dissociative experiences. When these measures present values below or above certain thresholds, they describe the presence of character and temperament fragility and of abnormal reactions to traumatic events. Specifically, scores above 70° percentiles as per the TCI temperamental scales, described as "Novelty Seeking" (NS) and "Harm Avoidance" and scores below 30° percentiles as per the character scales described as "Self Directiveness" (SD) and "Cooperativeness" (Coop) have been correlated to the presence of BPD in the same way as the high scores have been correlated to the taxonomic factor of the DES (DEStax) (Meares, 2014).

The scores obtained in these specific scales, from subjects belonging to the sample being examined, were transformed in continuous values comprised between −1 and 1 with a procedure known as "feature scaling." Adjusted version of subtracting the minimum and dividing by the range $(2*(x - min)/(max - min))$ −1. Adjusted normalized values fall between −1 and 1. These values represent the input signals conducted by the nodes of the first layer, which, for this purpose, were labeled with the names of the 5 scales described above NS, HA, SD, Coop e DEStax).
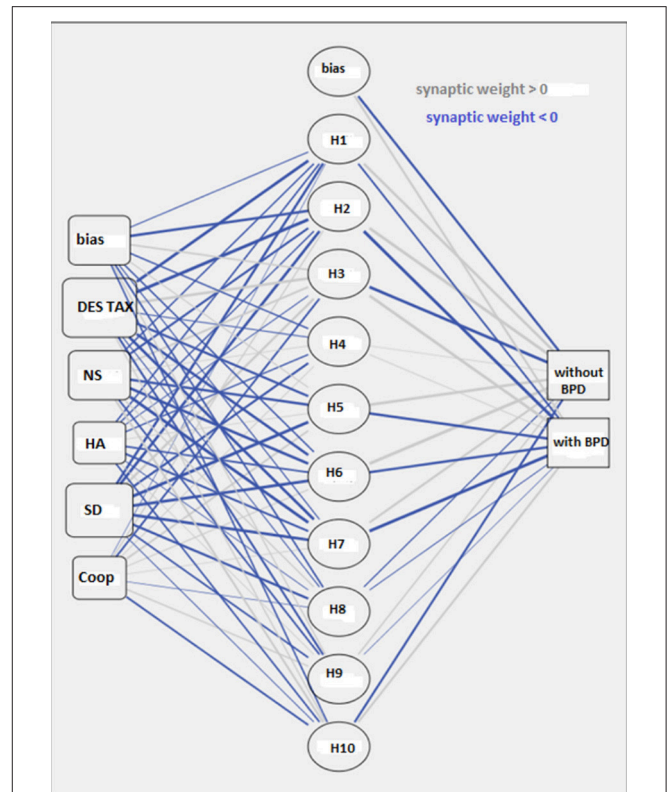
FIGURE 1 | Graph of the multilayer perceptron network.

Each of the 10 hidden layer computing units labeled with the letter H and an order number between 1 and 10 received, as an input, the scores of each of the 5 variables of the first layer. Each H node has produced two output values one for each of the two nodes of the third layer (Y), these nodes process information regarding the presence of BPD and are therefore labeled "with BPD" and "without BPD."

Each value of the neurons H and of the neurons Y has been associated with a weight "w" which is a multiplicative factor applied to the input signal; each node is associated with a bias "wb." The weighed sum of all the input signals of the nodes (H and Y) defines the internal activation of node named "A."

For example, in the case of the third output layer node "with BPD" the internal activation "A" is described by the following formula

Ay, withBPD =   H1w1 + H2w2.... + H10w10 + wby, withBPD

In the case of the node of the second layer "H1" the internal activation "A" is described by the following formula

AH1 = DEStaxwdestax + NSwns + HAwha + SDwsd + Coopwcoop + wbH1

The output signal of the nodes H and Y, defined as Neuronal Activity "AN," was calculated by applying a sigmoid-shaped transfer function

AN = 1/1 + e-A

Where "e" is the basis of the natural logarithms (2.71828) and A and the internal activity of the node.

For network training (**Figure 2**), the synapse weights were set to small random values; a training sample of 80% of the total sample was defined; the scores obtained from this subgroup of subjects to the 5 variables of TCI and DES were used to provide input data to the network; as output values coupled to input variables, the presence (with BPD) or absence (without BPD) of Borderline Personality Disorder Diagnosis was used. The core of the learning algorithm EBP is represented by a method of minimizing the error defined as "steepest descendent method" (Courant, 1943).

In the Back-Propagation phase, a "Batch" type weight correction process was applied in which all corrections are made at the end of all subgroup training cycles by using all the information of the training subjects. A repetition of the correction process was planned so as to obtain an effective output that did not deviate by more than 0.33 for each sample from the sample output used for training.

Each weight has been altered by an amount equal to the delta of the neuron multiplied by the input value on that connection, which is multiplied by an Epsilon (ϵ) factor, labeled learning factor equal to 0.1 selected to obtain a high number of epochs of learning which would guarantee a more accurate model and avoid the risk of a local minimum disabling the process.

Training of the Intermediate neurons was done by back-propagation through the delta error that had been calculated on the nodes of the third layer.

Finally, because the connection proceeds to several neurons of the next layer the error was calculated as the sum of all the individual Δ weighed. Calculation of the learning and verification Phase, model description, discrimination analysis, ROC Curves and analysis of the importance of the variables were performed with the help of the Statistical Package for Social Science (SPSS) software. This computer instrument, even assuming the initial weight values in random order, is able to repeat the data by setting the random number generator to an initial fixed value.

ANALYSIS OF THE RESULTS

Figure 3 reviews the summary of the regression model produced by the multilayer perceptron network where the percentage of correctness of the prediction of the presence and absence of the BPD are assessed; the model appears valid in terms of statistical reliability.

The learning process developed on a group of 480 subjects which represent 80% of the total sample (described at the top of the table) stopped after less than 100 cycles because it reached an error level below the selected value of 0.33. This value was chosen because the approximation of less than 33% to the output value 1 that expresses BPD is sufficient to guarantee the accuracy of the forecast. This value of tolerated difference between the desired output and the output observed in the single case allows us to calculate the accepted value of the error of sum of squares according to the following formula:

$$E = (1/2)KNe$$

where E = error of the sum of squares, K = number of cases used for learning, N = number of output nodes, e = tolerated difference between the desired value and the value observed in the individual case (Pessa, 2004).

$$(1/2) \cdot 480 \cdot 2 \cdot 0.332 = 52$$

The error of the sum of squares has reached a value of 46.77 at the end of the training phase. The percentage of incorrect predictions in the learning sample was 16.5%. The test sample (described in the lower part of **Figure 3**) is made up of 120 subjects. In this sample, the sum of squares has reached a value of 8.19 and the percentage of incorrect predictions was 10.1%.

Figure 4 shows the ROC curves for the sensitivity and specificity of the model in predicting the presence and an absence

Input layer – 5 continuous variables	
DES TAX	DES Taxonomic factor
NS	Novelty Seeking
HA	Harm Avoidance
SD	Self Directiveness
Coop	Cooperativeness
Scale of the variables	Normalizzata adattata
Number of hidden layers (H)	1
Number of units in the H layer	10
Activation function	Sigmoide
Output layer (Y)	DBP absent/present
Number of units	2
Activation function	Sigmoide
The value of the learning factor (ε)	0.1
Difference between the observed output and the desired output	0.33
Tolerated value of the error of the sum of squares	< 52

FIGURE 2 | Design of the multilayer perceptron network.

Training sample	Observed events	Predicted events		Percentage of correctness
		BPD absent	BPD present	
N. Of subjects	DBP absent	381	14	96,50%
480 (80%)	DBP present	65	20	23,50%
		92,90%	7,10%	83,50%
Precentage of mistaken predictions		16,50%		
Errors of the sum of the squares		46.77		

Test sample	Observed events	Predicted events		Percentage of correctness
		BPD absent	BPD present	
N. Of subjects	DBP absent	104	0	100,00%
N. 120 (20%)	DBP present	12	4	23,10%
		97,00%	3,00%	89,9%
Precentage of uncorrect predictions		10,10%		
Errors of the sum of the squares		8.19		
Dependent variable: BPD present/absent				

FIGURE 3 | Summary of the model and discriminant analysis.

FIGURE 4 | Curves receiver operating characteristic (ROC).

FIGURE 5 | Importance of the independent variables.
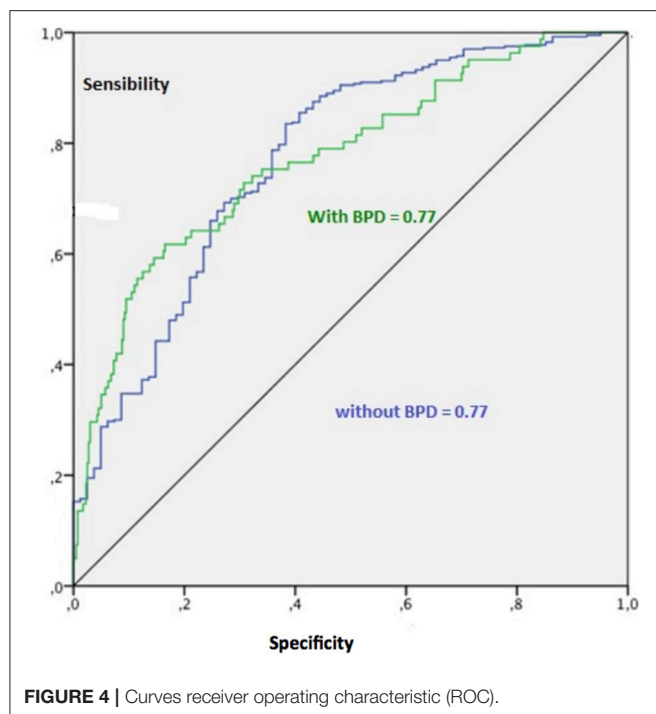
of diagnosis of BPD. The value of the area under the curve of 0.77 expresses a good sensitivity of the model. **Figure 5** shows the analysis of the importance of the variables (Towell and Shavlik, 1992), which shows that high scores at the taxonomic scale of DES and Novelty Seeking scale have a weight of 40 and 20% respectively in the model in question, while Self-Directiveness scales have a weight of 28.5% in determining the model's prediction accuracy.

DISCUSSION

The study data describes a non-linear relationship between the variables in question and show how in the genesis of BPD, the presence of dissociative phenomena has a double weight of character and temperament variables (SD and NS). DES is an instrument that detects dissociative phenomena rather than their pathological nature. The taxonomic scale (DEStax) correlates more than the other scales with dissociative syndromes classified in DSM. However, it does not consent to formulate any diagnosis.

Subjects with high scores at DEStax are not necessarily affected by dissociative disorders and often exhibit a mode that is biologically determined to respond to traumatic events that that they seem to manipulate by means of detachment, amnesia or imaginative absorption. From our study, it can be seen that when these subjects, who have an anomalous reactivity to trauma, have a marked explorative tendency or poor ability in self-determination functions because of their life experiences, there is a 90% probability that they can develop a BPD.
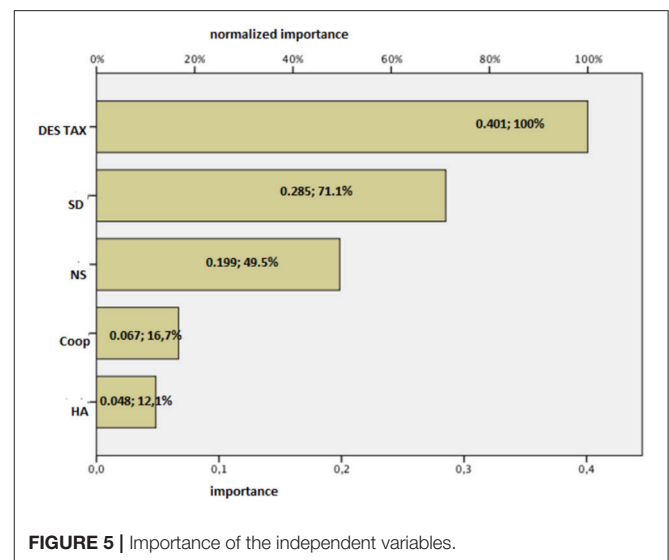
As proposed by Borsboom, the pathology develops on the basis of an interaction between psychopathological processes; he specifies that this interaction concerns the individual symptomatic level so the nodes of the network are internal to the pathology itself (Jones et al., 2017).

Our study said that, at least as regards personality disorders (Sperandeo et al., 2016), nonlinear interaction is also about pre-pathological phenomena such as extreme expressions of dissociative dimensions and personality, as many psychic phenomena as possible and environmental events. This integrative vision also emerges also in the commentary by Jones et al., on the work of Borsboom. These authors propose "an expansion of the network theory of psychopathology in which nodes consist of individual level causal variables." Expanding the network approach beyond symptoms will further strengthen this potentially revolutionary framework for studying psychopathology. These evidences confirm the validity of the intuitions of Borsboom with about the utility of imagining a psychopathological theory of networks. However, in our opinion it is important to integrate descriptive elements of psychopathological functioning and environmental stressors into this theoretical vision.

The dimensions of Temperament and Character detected by the TCI and the dissociative experiences described by DES are phenomenological, methodologically and theoretically different descriptors from the SCID items. In fact, they do not describe symptoms but ordinary processes that, on the one hand do not activate defense mechanisms in users (who tend to reject symptomatic stigmatization) and on the other hand are detectable before the symptoms actually arise.

The usefulness of a model that approaches the development of BPD starting from pre-symptomatic psychic processes is evident in the possibility of creating predictive tools for the

future development of the disease. Knowledge of the form of the interaction between personality dimensions and life experiences is essential to implement secondary prevention interventions.

It is necessary to apply the model to larger test samples to confirm the predictive validity of the same; subsequently it will be possible to integrate it with descriptive dimensions of the environmental experience and of the more refined and specific character processes in order to derive the description of the etiopathogenetic interaction from the model. In the future perspective we intend to integrate the predictors of the pathology obtaining a complete pattern psychical and environmental conditions significant for the pathology is to activate follow-up studies to test the model's ability to predict the development of disorders.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of codice etico per la ricerca in psicologia, comitato etico Associazione Italiana di Psicologia with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the comitato etico per la ricerca della SiPGI Postgraduate School in Gestalt Integrated Psychotherapy D.M.I.U.R. 12.10.2007.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Ball, J. S., and Links, P. S. (2009). Borderline personality disorder and childhood trauma: evidence for a causal relationship. *Curr. Psychiatry Rep.* 11, 63–68. doi: 10.1007/s11920-009-0010-4

Bernstein, E. M., and Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *J. Nerv. Mental Dis.* 174, 727–735.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry* 16, 5–13. doi: 10.1002/wps.20375

Borsboom, D., and Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Brodsky, B. S., Cloitre, M., and Dulit, R. A. (1995). Relationship of dissociation to self-mutilation and childhood abuse in borderline personality disorder. *Am. J. Psychiatry* 152, 1788–1792. doi: 10.1176/ajp.152.12. 1788

Cantone, D., Sperandeo, R., Maldonato, M. N., Cozzolino, P., and Perris, F. (2012). Fenomeni dissociativi in un campione di pazienti ambulatoriali. *Riv. Psichiatr.* 47, 246–253. doi: 10.1708/1128.12448

Carlier, I. V., Lamberts, R. D., Fouwels, A. J., and Gersons, B. P. (1996). PTSD in relation to dissociation in traumatized police officers. *Am. J. Psychiatry* 153, 1325–1328 doi: 10.1176/ajp.153.10.1325

Cloninger, C. R. (2000). A practical way to diagnosis personality disorder: a proposal. *J. Pers. Disord.* 14, 99–108. doi: 10.1521/pedi.2000.14. 2.99

Cloninger, C. R., Svrakic, D. M., and Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Arch. Gen. Psychiatry* 50, 975–990.

Courant, R. (1943). Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Am. Math. Soc.* 49, 1–23. doi: 10.1090/S0002-9904-1943-07818-4

Fabbri Bombi, A., Bertin, I., Cristante, F., and Colombo, F. (1996). Un contributo alla standardizzazione della Dissociative Experiences Scale (DES) di Bernstein e Putnam. *Bollettino di Psicologia Applicata* 219, 39–46.

First, M. B., Gibbon, M., Spitzer, R. L., and Benjamin, L. S. (1997). *User's Guide for the Structured Clinical Interview for DSM-IV Axis II Personality Disorders: SCID-II.* Washington, DC: American Psychiatric Pub.

Grabe, H. J., Spitzer, C., and Juergen Freyberger, H., (1999). Relationship of dissociation to temperament and character in men and women. *Am. J. Psychiatry* 156, 1811–1813. doi: 10.1176/ajp.156.11.1811

Jones, P. J., Heeren, A., and McNally, R. J. (2017). Commentary: a network theory of mental disorders. *Front. Psychol.* 8:1305. doi: 10.3389/fpsyg.2017.01305

Joyce, P. R., Light, K. J., Rowe, S. L., Cloninger, C. R., and Kennedy, M. A. (2010). Self-mutilation and suicide attempts: relationships to bipolar disorder, borderline personality disorder, temperament and character. *Aust. N. Z. J. Psychiatry* 44, 250–257. doi: 10.3109/00048670903487159

Kendler, K. S. (2012). The dappled nature of causes of psychiatric illness: Replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Mol. Psychiatry* 17, 377–388. doi: 10.1038/mp. 2011.182

Korzekwa, M. I., Dell, P. F., and Pain, C. (2009). Dissociation and borderline personality disorder: an update for clinicians. *Curr. Psychiatry Rep.* 11, 82–88. doi: 10.1007/s11920-009-0013-1

Maffei, C., Fossati, A., Agostoni, I., Barraco, A., Bagnato, M., Deborah, D., et al. (1997). Interrater reliability and internal consistency of the structured clinical interview for DSM-IV axis II personality disorders (SCID-II), version 2.0. *J. Pers. Disord.* 11, 279–284. doi: 10.1521/pedi.1997.11.3.279

Maldonato, N. M. (2009). From neuron to consciousness: for an experience-based neuroscience. *World Futures* 65, 80–93. doi: 10.1080/026040208025 94857

Maldonato, M. (2014). "The ascending reticular activating system," in *Recent Advances of Neural Network Models and Applications. Smart Innovation, Systems and Technologies, Vol. 26*, eds S. Bassis, A. Esposito, and F. Morabito (Cham: Springer), 333–344.

Maldonato, N. M., Dell'Orco, S., and Springer, M. (2011). "Rethinking consciousness: some hypothesis on the role of the ascending reticular activating system in the global workspace," in *WIRN* (Amsterdam), 212–219.

Meares, R. (2014). *Un Modello Dissociativo del Disturbo Borderline di Personalità.* Milano: Raffaello Cortina Editore.

Pessa, E. (2004). *Statistica con le Reti Neurali: un'Introduzione.* Roma: Di Renzo.

American Psychiatric Association (1996). *DSM-IV Manuale Diagnostico e Statistico dei Disturbi Mentali.* Milano: Ed. Masson.

Pukrop, R. (2002). Dimensional personality profiles of borderline personality disorder in comparison with other personality disorders and healthy controls. *J. Pers. Disord.* 16, 135–147. doi: 10.1521/pedi.16.2.135. 22550

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). *Learning Internal Representations by Error Propagation (No. ICS-8506).* San Diego, CA: California University San Diego, La Jolla Institute for Cognitive Science. doi: 10.21236/ADA164453

Sperandeo, R., Esposito, A., Maldonato, M., and Dell'Orco, S. (2016). "Analyzing correlations between personality disorders and frontal functions: a pilot study," in *Advances in Neural Networks. Smart Innovation, Systems and Technologies, Vol. 54*, eds S. Bassis, A. Esposito, F. Morabito, and E. Pasero (Cham: Springer), 293–302.

Sperandeo, R., Monda, V., Messina, G., Carotenuto, M., Maldonato, N. M., Moretto, E., et al. (2018). Brain functional integration: an epidemiologic study on stress-producing dissociative phenomena. *Neuropsychiatr. Dis. Treat.* 14, 11–19. doi: 10.2147/NDT.S146250

Spiegel, D., Hunt, T., and Dondershine, H. E. (1988). Dissociation and hypnotizability in posttraumatic stress disorder. *Am. J. Psychiatry* 145, 301–305 doi: 10.1176/ajp.145.3.301

Spiegel, D., Loewenstein, R. J., Lewis-Fernández, R., Sar, V., Simeon, D., Vermetten, E., et al. (2011). Dissociative disorders in DSM-5. *Depress. Anxiety* 28, E17–E45. doi: 10.1002/da.20923

Svrakic, D. M., Whitehead, C., Przybeck, T. R., and Cloninger, C. R. (1993). Differential diagnosis of personality disorders by the seven-factor model of temperament and character. *Arch. Gen. Psychiatry* 50, 991–999. doi: 10.1001/archpsyc.1993.01820240075009

Towell, G., and Shavlik, J. W. (1992). "Interpretation of artificial neural networks: mapping knowledge-based neural networks into rules," in *Advances in Neural Information Processing Systems* (San Mateo, CA), 977–984.

Waller, N., Putnam, F. W., and Carlson, E. B. (1996). Types of dissociation and dissociative types: a taxometric analysis of dissociative experiences. *Psychol. Methods* 1, 300–321. doi: 10.1037/1082-989X. 1.3.300

Yager, J. (1976). Psychiatrically maladjusting soldiers. *Arch. Gen. Psychiatry* 33, 1332–1335. doi: 10.1001/archpsyc.1976.01770110060005

Zlotnick, C., Shea, M. T., Pearlstein, T., Simpson, E., Costello, E., and Begin, A. (1996). The relationship between dissociative symptoms, alexithymia, impulsivity, sexual abuse, and self-mutilation. *Compr. Psychiatry* 37, 12–16. doi: 10.1016/S0010-440X(96)90044-9

Validation of a Spanish Questionnaire on Mobile Phone Abuse

María A. Olivencia-Carrión[1], Isabel Ramírez-Uclés[2], Pablo Holgado-Tello[3] and Francisca López-Torrecillas[1]*

[1] Center for Research into the Mind, Brain and Behavior, Granada University, Granada, Spain, [2] Department of Personality, Assessment and Psychological Treatment, Universidad Nacional de Educación a Distancia, Madrid, Spain, [3] Department of Behavioral Sciences Methodology, Universidad Nacional de Educación a Distancia, Madrid, Spain

Mobile phone addiction has attracted much attention recently and is showing similarity to other substance use disorders. Because no studies on mobile phone addiction had yet been conducted in Spain, we developed and validated a questionnaire (Cuestionario de Abuso del Teléfono Móvil, ATeMo) to measure mobile phone abuse among young adults in Spanish. The ATeMo questionnaire was designed based on relevant DSM-5 diagnostic criteria and included craving as a diagnostic symptom. Using stratified sampling, the ATeMo questionnaire was administered to 856 students (mean age 21, 62% women). The MULTICAGE questionnaire was administered to assess history of drug abuse and addiction. Using confirmatory factor analysis, we found evidence for the construct validity of the following factors: Craving, Loss of Control, Negative Life Consequences, and Withdrawal Syndrome, and their association with a second order factor related to mobile phone abuse. The four ATeMO factors were also associated with alcoholism, internet use, and compulsive buying. Important gender differences were found that should be considered when studying mobile phone addictions. The ATeMo is a valid and reliable instrument that can be used in further research on mobile phone abuse.

Keywords: mobile phone, DSM-5, validity, Spanish population, abuse

INTRODUCTION

The mobile phone has many characteristics that make it attractive to young adults. It is primarily used to communicate but also has many other interesting applications, including camera, internet, music reproduction, games, and social media. The International Telecommunication Union report (The International Telecommunication Union, 2016) finds that 98% of young adults own a mobile phone in Europe and other studies indicate that young women in particular have more interest in mobile phones than other groups do (Roberts et al., 2014). There is evidence that mobile phone abuse in related to physical and mental wellbeing problems, including social and psychological disturbances such as attention deficit and hyperactivity disorder, disruptive behavior disorders, anxiety disorders, mood disorders, substance use disorders, sleep disorders, and eating disorders (Billieux et al., 2014; Foerster et al., 2015). In recent years, a co-occurrence has been established between mobile phone dependence and other behavioral disorders such as internet addiction (Chiu et al., 2013), compulsive buying (Jiang and Shi, 2016) and alcohol use (De-Sola et al., 2017a) or use of other substances (Gallimberti et al., 2016). However, it remains unclear if an individual that develops one addictive behavior (i.e., mobile phone abuse) is more likely to develop another addictive behavior or a substance use problem.

Although a definition of mobile phone abuse has not yet been agreed upon, some researchers define mobile phone dependence as a constant use of the device with a poor capacity to control daily activities, exhibiting extreme nervousness and aggressive behavior when deprived of its use; this excessive use is also accompanied by a progressive deterioration in school/work performance and social and family functioning (Billieux et al., 2014; Lin et al., 2015). These symptoms have a major negative impact on the life of the affected person, reflected in impaired health or deprived social functioning; they have also been shown to be equivalent to substance dependence as understood by the current nosological systems Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5, American Psychiatric Association, 2012).

Mobile phone addiction could in many ways be similar to substance dependence disorders (Foerster et al., 2015; Roser et al., 2016). For instance, the abuse of psychotropic drugs (heroin, cocaine, cannabis, etc.) and alcohol is a complex social, biological, and psychological phenomenon. Whether an individual ever uses alcohol or another substance, and whether that initial use progresses to a substance use disorder of any severity, depends on a number of factors. These include: a person's genetic makeup and other individual biological factors; psychological factors related to a person's unique history and personality; and environmental factors, such as the availability of drugs, family and peer dynamics, coping with stress, and access to social support. Chronic consumption of several drugs (cannabis, stimulants, and opioids) has been associated with the presence of neuropsychological impairments in a broad range of functions. In recent years neuropsychological research on substance abuse has focused on the study of impairments in executive functions linked to the prefrontal cortex and their influence on the personality, cognitions, and behaviors of the substance abusers (López-Torrecillas et al., 2000; Verdejo-García et al., 2004).

To date, pathological gambling is the non-substance related addiction which has received most attention and has been examined extensively. The results reveal a number of substantial similarities between pathological gambling and substance-related addictions concerning phenomenology, epidemiology, personality factors, genetics, neurobiological processes, recovery, and treatment (Walther et al., 2012; Contreras-Rodríguez et al., 2016; Navas et al., 2017). In DSM-5, pathological gambling is classified as a non-substance-related addiction and is, therefore, removed from the former category "Impulse-Control Disorders" and included in the new "Substance Use and Addictive Disorders" category. Other potential non-substance-related addictions are internet addiction, compulsive buying, sex addiction, and mobile phone addiction, although these are not yet officially defined as disorders due to a lack of evidence. Despite a substantial overlap, it is not yet clear why some people become vulnerable to these behaviors. The co-occurrence of non-substance-related addiction with different forms of substance abuse such as smoking, drinking, use of cannabis, and other illegal drugs among young people has been repeatedly discussed (Vanyukov et al., 2012; De-Sola et al., 2017a).

The literature also reveals an association between multiple substance use and other risk behaviors among young adults. For example, binge drinking, cannabis use, and tobacco use appear to be more prevalent in young people (Van Rooij et al., 2014; Abebe et al., 2015). The use of both alcohol and cannabis predicts use of common addictive substances (Osuch et al., 2013; Viola et al., 2014; Vorspan et al., 2015) and tends to be accompanied by gambling (Larsen et al., 2013; Míguez and Becoña, 2015). In addition, a number of authors (Mudry et al., 2011; Yau et al., 2012; Grant et al., 2013; Lee et al., 2013; Mattebo et al., 2013; Schuster et al., 2013; Van Rooij et al., 2014; Biolcati, 2015) have pointed out the relationship between the amount of time young adults spend gambling, abusing their mobile phones, using the internet, playing video games, buying compulsively, or having sex and increases in alcohol, tobacco, cannabis, and drug consumption.

The acknowledgement of behavioral addictions as disorders can be traced as far back as Marlatt et al. (1988), who reported a repetitive habit pattern that increases the risk of disease and/or associated personal and social problems. Addictive behaviors are often experienced subjectively as a loss of control in which the behavior continues to occur despite volitional attempts to abstain or moderate use. Furthermore, in the last decade, a growing number of studies have established psychological and neurobiological similarities between the excessive practice of addictive behaviors (e.g., mobile phone abuse, compulsive buying, sex, internet, video gaming, and eating disorders; Billieux et al., 2010; Mentzoni et al., 2011). Research on the neurobiology of addiction has also found a common mechanism between substance addictions and behavioral addictions (Leeman and Potenza, 2013; Weinstein and Lejoyeux, 2015). However, at this point we do not know whether having one addictive behavior increases the likelihood of developing other addictive behaviors or other dependencies such as substance use disorders. In addition, alcohol, drugs, and pathological gambling may not be the only crippling addictions that we should address. Unfortunately, other addiction statistics are scarce because many destructive habits are not yet officially recognized as addictions, including mobile phone addiction, game addiction, eating, shopping, and sex addiction, all of which are problematic for many reasons. They all involve direct manipulation of pleasure through the use of products, similar to drug use disorders and food-related disorders.

The concept of non-substance-related (or "behavioral") addiction describes syndromes analogous to substance addiction, but with a focus on a certain behavior which, similar to substance consumption, produces short-term reward and may persist despite harmful consequences due to diminished control over the behavior. Given that addictive behavior is not necessarily restricted to substance consumption, the DSM-5 broadens the category "Substance-Related Disorders" to a "Substance Use and Addictive Disorders" category including both substance and non-substance-related addictions. The Diagnostic and Statistical Manual of Mental Disorders—4th Edition (DSM-IV; American Psychiatric Association , 2002) conceptualized two discrete substance use disorders (SUD), abuse and dependence, defined by mutually exclusive sets of diagnostic criteria. Abuse required endorsement of one or more (1+) of four abuse criteria, and dependence required endorsement of three or more (3+) of seven dependence criteria. In contrast, the proposed Diagnostic

and Statistical Manual of Mental Disorders—5th Edition (DSM-5; American Psychiatric Association, 2012) conceptualizes a unitary SUD construct, varying only in terms of severity. The literature reviewed here includes studies on postulated behavioral addictions related to the use of mobile phones, shopping, sex, internet, video gaming, and food, along with other studies that analyzed the co-occurrence of these addictions with substance abuse (for instance tobacco, alcohol, and cannabis substances). However, these are not included in the DSM-5 because of current lack of evidence. In order to be able to obtain relevant evidence in the first place, we need valid and reliable instruments that allow us to measure addictive behaviors such as mobile phone abuse.

The study of mobile phone abuse started in 2004 with the development of the Mobile Phone Dependency Questionnaire (CPDQ; Toda et al., 2004) designed for use in university populations and validated in a population of high school students by Kawasaki et al. (2006). Another instrument available for use in adult populations is the Mobile Phone Problem Use Scale (MPPUS; Bianchi and Phillips, 2005); including a recent short version (Foerster et al., 2015) and a version for teenagers [Mobile Phone Addiction Scale (MPAS; Leung, 2008)]. This scale has been translated into Japanese (Takao et al., 2009) and Spanish (López-Fernández et al., 2012), with some items previously translated for use in the Spanish university population (Ruiz-Olivares et al., 2010). Although the Mobile Phone Problem Use Scale is one of the most frequently used instruments to assess mobile phone addiction, other instruments exist including the Mobile Phone Usability Questionnaire (MPUQ; Ryu and Smith-Jackson, 2006) and Problematic Mobile Phone Use Questionnaire (PMPUQ; Billieux et al., 2008). In Eastern countries, three scales have been developed: the Mobile Phone Dependence Inventory (MPDI; Xu et al., 2008); the Excessive Cellular Phone Use Survey (ECPUS; Ha et al., 2008), and the Smartphone Addiction Scale (SAS-SV; Kwon et al., 2013). At present there are 5 instruments translated into Spanish: the first one, already mentioned previously—MPPUS, Bianchi and Phillips (2005)—has been adapted by López-Fernández et al. (2012); the second one is the Cell-phone Over-use Scale (COS; Jenaro et al., 2007) for university populations; the third one is the Questionnaire of Mobile-Related Experiences (CERM; Fargues et al., 2009) for adult populations; the fourth one is the Test for Mobile Phone Dependence [TMD] (Chóliz, 2012) for adolescents (including a new reduced version, Chóliz et al., 2016), and finally, the fifth is a questionnaire that focuses only on the dimension of Craving (De-Sola et al., 2017b).

However, no studies have been conducted in Spain to identify mobile phone addiction in young adults using the DSM-5 criteria (American Psychiatric Association, 2012). To this end, it is necessary to develop a valid and reliable instrument measuring mobile addiction, having in mind the modifications made in the DSM-5 (American Psychiatric Association, 2012). These modifications imply that mobile phone addiction should be considered in relation to substance use disorders and behavioral addictions. The diagnostic symptoms of substance use disorders since recently include a new criterion, *craving*, featured in the DSM-5 (American Psychiatric Association, 2012). One of the most accepted definitions of craving is that of compulsive craving—an irrational and intense desire or uncontrollable

compulsion to consume a particular psychoactive substance and/or perform a certain behavior, which leads to compulsive search rituals (Blasco et al., 2008; Igarashi et al., 2008; De-Sola et al., 2017b). Hence, craving should be considered as a criterion to establish a diagnosis and understand the different mediating variables when developing treatments, analyzing relapses, and designing prevention strategies. Therefore, the main purpose of this study was to develop and validate a questionnaire to measure mobile phone dependence among young Spanish speaking adults.

METHODS

Participants

The sample comprised of community-dwelling young adults between 17 and 45 years of age (mean 21.12 years old, standard deviation=3.05, 62.38% women and 37.62% men). They were recruited from the student population of the University of Granada. Participants were recruited by university faculty during class breaks and were selected using a probabilistic sampling design. In particular, a cluster stratified sample design was adopted. Strata were based on the different university faculties. Cluster samples were extracted such that majors and years of study were represented in proportion to the total number of students in each faculty. Finally, all students of the cluster sample were included in the final sample. There were 856 participants recruited between September 2013 and June 2014. The participants were informed about the aims of the study and provided signed informed consent prior to participation. Inclusion criteria were having a mobile phone, wanting to participate, and signing the informed consent form. Prior to recruitment the study was approved by the Research Ethics Committee from the University of Granada, Spain.

Measures

The Mobile Phone Abuse Questionnaire (ATeMo)

The Mobile Phone Abuse Questionnaire (ATeMo) was developed to assess mobile phone dependence. It consists of 25 items covering addictive symptoms, based on the diagnostic criteria for behavioral addiction (gambling) and the DSM-5 (American Psychiatric Association, 2012), and also taking into account substance abuse disorders, and instruments that measure addiction to mobile phones, internet, and social networks. The addictive symptoms considered were craving, loss of control, negative life consequences, and withdrawal syndrome. Specifically, the questionnaire assessed the use of the mobile phone, the disturbance of daily activities, the increase in time spent to obtain the same satisfaction, loss of control, difficulties in stopping using the phone and the irritability produced, and the negative feelings experienced when the mobile phone cannot be used. The 25 items were answered on a 5-point Likert scale that ranged from 0 (strongly disagree) to 4 (strongly agree), resulting in a final score between 0 and 100 (see **Table 1**).

The MULTICAGE CAD-4

The MULTICAGE CAD-4 was designed to screen for a history of drug abuse and addiction behavior. It assesses alcoholism (items

TABLE 1 | Mobile Phone Abuse Questionnaire (ATeMo).

1. When I forget my mobile phone I feel restless.

   (*Cuando me olvido del móvil me siento intranquilo*)

2. I'd rather lose my wallet than my mobile phone.

   (*Prefiero perder la cartera que el móvil*)

3. I don't want to go to places where the mobile signal is weak.

   (*No quiero ir a lugares donde la señal de móvil sea débil*)

4. When I travel I often touch my mobile phone.

   (*Cuando viajo suelo estar tocando el móvil*)

5. I use whatsapp/line or similar more than 4 hours a day.

   (*Uso whatsapp/line o similar más de 4 horas al día*)

6. I use whatsapp/line or similar while I study/work.

   (*Uso whatsapp/line o similar mientras estudio/trabajo*)

7. I use whatsapp/line or similar when I'm with my friends or my family.

   (*Uso whatsapp/line o similar cuando estoy con amigos o familia*)

8. I use whatsapp/line or similar at night, in bed before going to sleep.

   (*Uso whatsapp/line o similar por la noche, en la cama antes de dormir*)

9. I unconsciously check whatsapp or the messages I have.

   (*Inconscientemente compruebo los whatsapp o mensajes que tengo*)

10. I feel happy when I receive a message or whatsapp.

    (*Me siento feliz cuando recibo un mensaje o whatsapp*)

11. I express my feelings better through whatsapp than talking.

    (*Expreso mis sentimientos mejor con los whatsapp que hablando*)

12. I can never spend enough time on my mobile phone.

    (*Nunca puedo estar el tiempo que necesito utilizando mi móvil*)

13. I have used my mobile phone to make myself feel better when I was feeling down.

    (*Utilizo el móvil para sentirme mejor cuando estoy bajo de ánimo*)

14. I lose sleep due to the time I spend on my mobile phone.

    (*Pierdo horas de sueño utilizando el móvil*)

15. When out of range for some time, I become preoccupied about the thought of missing a whatsapp or message.

    (*Cuando estoy sin cobertura me preocupo por los whatsapp o mensajes que pueda perder*)

16. I have attempted to spend less time on my mobile phone but I'm unable to.

    (*He intentado gastar menos tiempo usando el móvil pero no lo consigo*)

17. I have aches and pains that are associated with my mobile phone use.

    (*Tengo molestias físicas o dolores producidos por el uso del móvil*)

18. I become irritable if I have to switch off my mobile phone for a meeting, dinner engagements or when at the movies.

    (*Me irrita tener que apagar el móvil en situaciones como reuniones, cenas, encuentros o en el cine*)

19. I feel lost without my mobile phone.

    (*Me siento perdido sin mi teléfono móvil*)

20. I feel angry if someone interrupts me when I'm using my mobile phone.

    (*Me enfado si alguien me interrumpe cuando estoy usando móvil*)

21. If I had to spend 6 h without using my cell phone, I would feel restless or nervous.

    (*Si tuviera que pasar 6 horas sin utilizar el móvil, me sentiría inquieto o nervioso*)

22. I feel bored when I'm not using my mobile phone.

    (*Me siento aburrido cuando no estoy utilizando el móvil*)

23. I neglect my work or class assignments to use my mobile phone.

    (*Desatiendo mis tareas de trabajo o clase para usar el móvil*)

(Continued)

TABLE 1 | Continued

24. I ignore my friends to use my mobile phone.

    (*Desatiendo a mis amigos para usar el móvil*)

25. I ignore my family to use my mobile phone.

    (*Desatiendo a mi familia para usar el móvil*)

Instructions: We are interested in how people use mobile phones to communicate. Please indicate the degree to which you agree or disagree with each of the following statements regarding your use of your mobile phone on the following scale:
0, strongly disagree; 1, disagree; 2, neutral; 3, agree; and 4, strongly agree.
Scoring and interpretation: Sum the items in parenthesis for subscale scoring: Craving (1, 2, 3, 4, 9, 10, 12, 16), Loss of Control (5, 8, 13, 14), Negative Life Consequences (6, 7, 11, 17, 18, 23, 24, 25) and Withdrawal Syndrome (15, 19, 20, 21, 22).

1–4), gambling disorders (items 5–8), drug addiction (items 9–12), eating disorders (items 13–16), internet addiction (items 17–20), video gaming addiction (items 21–24), compulsive buying disorder (items 25–28) and sex addiction (items 29–32). The psychometric properties have been well established in Spanish adult populations. It demonstrates high > 0.7 Cronbach's alpha coefficient. In the exploratory factor analysis, 8 components are identified that identify the proposed structure the diagnostic sensitivity for alcohol was 92.4%, and between 94 and 100% for heroin, cocaine and cannabis (Pedrero-Pérez et al., 2007).

Procedure

The study consisted of two stages: in the first stage the instrument was developed and in the second stage it was validated. The construction of the Mobile Phone Abuse Questionnaire (ATeMo) was based on the DSM-5 (American Psychiatric Association, 2012) that does not recognize mobile addiction as a disorder but makes reference to tobacco addiction and gambling. Ideas were taken from instruments that measure addictions to mobile phones, internet, and social networks and items were created taking into account all the aforementioned. For the construction of the items, criteria for constructing items for Likert questionnaires were used (Jenaro et al., 2007; Billieux et al., 2008; Fargues et al., 2009; Chóliz, 2012; Chóliz et al., 2016; De-Sola et al., 2017a). This set of defined criteria together with the items that evaluated them were reviewed by three experts on clinical psychology, educational psychology, and psychometrics. The experts collaborated in writing and ensuring the understanding, clarity, and consistency in the definitions of the criteria and the items. For the evaluation of the items a 5-point rating system was applied (from 0 to 4) taking into account the frequency from never to always (Fishman and Galguera, 2003; Schepers, 2009; Furr, 2011; DeVellis, 2012). Once the expert evaluation was concluded, a pilot experiment was carried out on a sample of 65 university students. They were asked to indicate whether the items in the questionnaire were comprehensible or not, encouraging them to raise any doubts that they had regarding each item.

The instrument was then administered to the final sample of participants in order to establish its validity. The data were collected from students of the University of Granada through stratified sampling by conglomerates, according to majors and groups of the different degrees taught at the

University of Granada (Psychology, Speech Therapy, Tourism, English, History, Literature, GADE, Economy, Biology, Physics, Optics, Primary, Infant, Pedagogy, Law, Medicine, Pharmacy, Social Work, Policies, Sociology, Information Technology, Roads and Telecommunications). Teachers responsible for the selected groups were sent an email informing them of the objectives of the study and requesting their help so that the students could participate. They were asked to inform their students about the study and the time during breaks was used to complete the questionnaires. It was emphasized that the participation was voluntary, that is, the students were free not to participate if they preferred it. The teachers also emphasized the need for honesty when filling out the survey and guaranteed the confidentiality of the responses. The survey started with short demographic questions (sex and age) followed by the ATeMo questionnaire.

Data Analysis

To obtain empirical evidence about the construct validity of the questionnaire and given the ordinal nature of the data, we conducted Confirmatory Factor Analysis (CFA) using polychoric correlations and Unweighted Least Squares (ULS) as estimation method (Hernández et al., 2000; Yang-Wallentin et al., 2010; Morata-Ramírez et al., 2015). We also tested the basic psychometric properties of the dimensions obtained (mean, standard deviation, reliability and discrimination). For criterion validity, correlation analysis was performed to determine the relationship of the ATeMo factors and the sub-dimensions of the MULTICAGE CAD-4. Gender differences between the different factors of the questionnaire were also examined through a MANOVA. Finally, to achieve an initial approximate interpretation of the scores, we calculated the percentiles in the total sample and split them by gender. The statistical programs used were SPSS 15.0 for Windows and LISREL 8.71 (Jöreskog and Sörbom, 1996).

RESULTS

Confirmatory Factor Analysis (CFA)

In order to obtain empirical evidence about the adequateness of the postulated structure ofthe ATeMo questionnaire, a CFA was conducted. In line with the theoretical background, the dimensional structure considered implied a general second-order factor referring to mobile phone dependence and four first order factors. The four first order factors were the following: eight items contributed to the first factor of Craving (1, 2, 3, 4, 9, 10, 12, 16), four items to the second factor of Loss of Control (5, 8, 13, 14), eight items to the third factor of Negative Life Consequences (6, 7, 11, 17, 18, 23, 24, 25) and five items to the fourth factor Withdrawal Syndrome (15, 19, 20, 21, 22). For the model examined (**Figure 1**), the Global fit Indices were: $\chi^2 = 274.18$; $d.f. = 265$; $p = 0.34$. The value of the Root Mean Square Error of Approximation (RMSEA) was 0.021, with a 90% interval between 0.0 and 0.050. The Goodness of Fit Index (GFI) was 0.97, the Adjusted Goodness of Fit Index (AGFI) was 0.97, the comparative Fit Index (CFI) was 1, the Normed Fit Index (NFI) was 1 and the Standardized Root Mean Square Residual (SRMR) was 0.06. These data show that the fit values of the model

are appropriate. All the lambdas and gammas parameters were statistically significant.

Reliability

The reliability of ATeMo was assessed using Cronbach's alpha coefficients (**Table 3**) and the resulting values were: Total score 0.91; Craving factor 0.74; the Loss of Control factor 0.70; Negative Life Consequences factor 0.77; and for the Withdrawal Syndrome factor 0.77. In addition, we calculated descriptors for the ATeMo from the CFA (mean, standard deviation and mean discrimination of the items of each dimension: **Table 2**).

Criterion Validity

To determine the criterion validity, we calculated the Pearson bivariate correlation index between the total score and each of the ATeMo factors, as well as with the MULTICAGE CAD-4 subscales (see **Table 3**). There was a positive correlation between the ATeMo total score and Alcoholism, Gambling disorders, Internet addiction, and Compulsive buying in the MULTICAGE CAD-4 subscales. Furthermore, there was a positive correlation between the Negative Life Consequences factor of ATeMo and Drug addiction in the MULTICAGE CAD-4 subscale; the Craving and Loss of Control ATeMo factors and Video game addiction in the MULTICAGE CAD-4 subscale; and the Negative Life Consequences and Withdrawal Syndrome factors of ATeMo and Sex addiction in the MULTICAGE CAD-4 subscale.

In general terms, the direction of the correlations is consistent with what was expected, however, given the large sample size, correlations of 0.073 (Craving-Video games addiction), for example, result statistically significant. For this reason, according to Rosnow and Rosenthal (1996) none of the correlations presents a large effect size ($|r| > 0.37$). The effect size is medium ($|r| > 0.24$) for the correlations between the factors of ATeMO with Internet Addiction, and Compulsive Buying. The correlation between the factors of ATeMO and Alcoholism and Eating Disorders have a low effect size ($|r| > 0.10$). All other correlations have an irrelevant effect size.

Score Interpretation

In order to provide preliminary data to help interpret the scores obtained, the 10th to 90th percentiles are presented for the total sample, and for men and women separately (**Table 4**).

DISCUSSION

In the present study we have developed a new valid and reliable scale to measure mobile phone abuse and dependence in Spain (ATeMo). The ATeMo Questionnaire consists of 25 items covering addictive symptoms, based on the diagnostic criteria of the DSM-5 (American Psychiatric Association, 2012). It is evaluated on a 5-point Likert-type scale ranging from 0 (strongly disagree) to 4 (agree), resulting in a final score in the range of 0–100. According to results from a confirmatory factor analysis, the ATeMo represents a general second order factor and four first order factors consistent with addiction theory: Craving, Loss of Control, Negative Life Consequences,
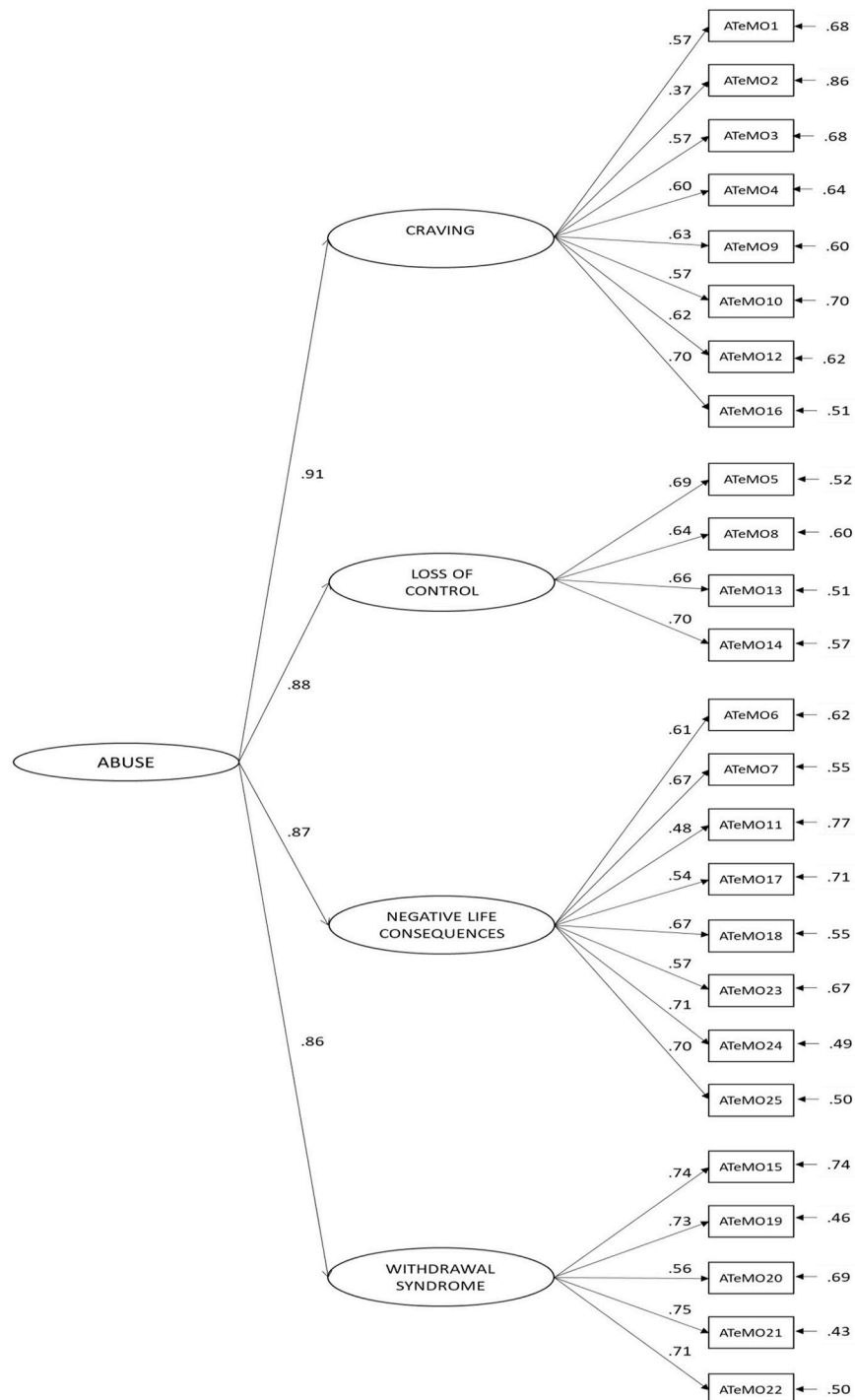
FIGURE 1 | $\chi^2 = 274.18$; d.f. = 265; $p = 0.34$; RMSEA = 0.021, GFI = 0.97; AGFI = 0.97; CFI = 1; NFI = 1; and SRMR = 0.06.

and Withdrawal Syndrome. These factors show considerable overlap with the symptoms proposed previously (Bianchi and Phillips, 2005; Rutland et al., 2007; Igarashi et al., 2008; Yen et al., 2009; Walsh et al., 2010; Chóliz, 2012; Merlo et al., 2013; Chóliz et al., 2016) and were developed according to the

criteria for the diagnostic symptoms of substance dependence disorders in the DSM-IV-TR (American Psychiatric Association, 2002) and the DSM-5 (American Psychiatric Association, 2012), the latter more recently including craving as a diagnostic criterion.

In assessing the reliability of the ATeMo questionnaire, Cronbach's alpha coefficients were calculated, demonstrating it had excellent internal consistency as seen elsewhere in similar studies in Spain (Chóliz, 2012; López-Fernández et al., 2012; Vanyukov et al., 2012; Chóliz et al., 2016). These coefficients were higher than those obtained in some previous studies (Fargues et al., 2009), where measures were developed according to the criteria for diagnosing symptoms of substance dependence disorders in DSM-IV-TR (American Psychiatric Association, 2002). The MULTICAGE CAD-4 subscales were used to determine potential criterion validity of ATeMo, identifying a positive correlation between the ATeMo total score and Alcoholism, Drug addiction, Eating disorders, Internet addiction, and Compulsive Buying subscales (Chiu et al., 2013; Gallimberti et al., 2016; Jiang and Shi, 2016; De-Sola et al., 2017a). Furthermore, there was a positive correlation between the Craving ATeMo factor, Alcoholism, Eating disorders, and Internet addiction, and a negative correlation with Video gaming addiction in the MULTICAGE CAD-4 subscale. Similarly, there was a positive correlation between the ATeMo factor Loss of Control and Alcoholism, Eating disorders, Internet addiction, and Compulsive buying, as well as a negative correlation with Gambling Disorders in the MULTICAGE CAD-4 subscale. This is consistent with the positive correlation between self-control and addiction identified previously (Jiang and Shi, 2016). Again, there was a positive correlation with Negative Life Consequences as an ATeMo factor and Alcoholism, Drug addiction, Eating disorders, Internet addiction, Compulsive buying, and Sex addiction in the MULTICAGE CAD-4 subscale, and there was a similarly positive correlation between Withdrawal Syndrome as an ATeMo factor and Alcoholism,

Eating disorder, Internet addiction, Video gaming addiction, and Compulsive buying. Indeed, loss of control, negative life consequences and withdrawal syndrome were already considered as diagnostic criteria for addiction disorders prior to DSM-5 (American Psychiatric Association, 2012).

The relationships described above are consistent with previous considerations that alcohol consumption may predict problematic mobile phone use (De-Sola et al., 2017a). They are also consistent with previous results on the relationship between Internet and mobile phone addiction (Chiu et al., 2013) and with previous results suggesting common impulsive aspects between compulsive buying and mobile phone addiction (Jiang and Shi, 2016).

Furthermore, the survey conducted indicated a common continuum of substance abuse and behavioral addictions, as identified previously in surveys that focused on such co-morbidity (Chiu et al., 2013; Jiang and Shi, 2016; De-Sola et al., 2017a; although an association between eating disorders and mobile phone abuse is yet to be found). These results suggest that alcohol, drugs, and pathological gambling may not be the only crippling addictions. Addiction statistics are scarce because many destructive habits (such as gaming, shopping, sex, etc.) are not yet officially recognized as addictions, although they could be problematic for many reasons. Some of these involve the direct manipulation of pleasure through the consumption of products like in the case of drug use disorders and food-related disorders.

The results obtained with ATeMo indicate that there are gender differences between males and females regarding mobile phone abuse, with scores ≥8 for the former and ≥10 for the latter potentially indicating mobile phone addiction. These results are consistent with previous findings indicating that females send more and longer texts, they talk for longer than men on the phone, and tend to regard mobile phones as a social tool (Roberts et al., 2014).

Our findings demonstrate that the ATeMo is a valid and reliable instrument that can be administered to different groups of university students. In addition, while this instrument was developed for university students, renewed construct validity and reliability analyses could convert it into a version suitable for adolescents.

Our results should be evaluated in view of several important limitations. First, the sample used in this study was relatively homogeneous with respect to age and educational level. Second,

TABLE 2 | Cronbach's alpha coefficients and the mean, standard deviation, and mean discrimination for the AteMo questionnaire derived from the CFA.

	F1 (Cv)	F2 (LC)	F3 (NLC)	F4 (WS)	ATeMo (TS)
Mean	11.85	6.71	7.48	4.02	30.07
Standard deviation	5.39	3.66	4.72	3.53	15.00
Cronbach's α	0.74	0.70	0.77	0.77	0.91
Mean discrimination	0.45	0.49	0.48	0.55	0.52

Cv, Craving; LC, Loss of Control; NLC, Negative Life Consequences; WS, Withdrawal Syndrome; and TS, Total Scores.

TABLE 3 | Correlations between the total score, the factors of ATeMo, and the MULTICAGE CAD-4 subscales.

	MULTICAGE CAD-4							
	1	2	3	4	5	6	7	8
F1(Cv)	0.127**	−0.001	0.013	0.210**	0.252**	−0.073*	0.287**	0.032
F2 (LC)	0.139**	−0.036	0.025	0.193**	0.218**	−0.075*	0.260**	012
F3(NLC)	0.196**	0.062	0.100*	0.172**	0.300**	0.037	0.305**	093**
F4 (WS)	0.163**	0.022	0.032	0.161**	0.293**	−0.047	0.278**	0.073*
ATeMO (TS)	0.180**	0.015	0.050	0.214**	0.307**	−0.044	0.328**	0.061

1, Alcoholism; 2, Gambling disorders; 3, Drug addiction; 4, Eating disorders; 5, Internet addiction; 6, Video games addiction; 7, Compulsive buying; 8, Sex addiction; F (Cv), Craving; F2 (LC), Loss of Control; F3 (NLC), Negative Life Consequences; F4 (WS), Withdrawal Syndrome; and ATeMO (TS), Total Scores. *p < 0.05, **p < 0.01, ***p < 0.001.

TABLE 4 | Percentiles, raw scores in ATeMo, mean and standard deviation for men and women.

Pc	Raw Scores		
	Men	Woman	Total
10	1	2	2
20	2	4	3
30	3	5	5
40	4	7	6
50	5	8	7
60	6	9	8
70	7	10	9
80	8	11	10
90	10	12	12
Mean	5.37	7.51	6.71
Sd.	3.46	3.54	3.66

Pc, percentile; Sd, standard deviation.

mobile phone addiction should be investigated in relation to a number of variables, such as demographic, personality, and clinical characteristics. This could advance our understanding of the interaction of humans with technology, as well as our understanding of the nature and causes of technology-related addictions. Overall, taking into account the lack of a valid and reliable questionnaire to measure the addiction to the mobile phone, ATeMo could be an adequate instrument to measure the mobile phone addiction in future investigations.

Regarding clinical implications, the development of the ATeMo questionnaire to detect mobile phone abuse is an important step in the development of diagnostic and treatment procedures and in the design of prevention and intervention strategies.

In future studies, it would be of interest to examine the problems associated with mobile phone use in relation to variables such as solitude, depression, self-esteem, well-being, academic success, and other demographic variables. Further studies into the problematic use of mobile phones will not only allow us to better understand this problem but they should provide information to aid the committees determining future DSM criteria, especially in relation to addictions associated with new technologies. Moreover, a more profound analysis thorough ROC curves of the cut-off thresholds should be performed to help interpret the scores obtained and to classify the subjects. Moreover, other construct validity evidences should be investigated. In this sense, invariance analysis by gender, of age group, for example, is necessary to obtain empirical evidences about the equivalence in the constructs and items operatized in ATeMO. Once guarantee this issue, Differential Item Functioning and a deep comparative analysis by the sorting variables considered will be necessary to ensure that the decisions made based on the test scores are valid.

In summary, we have developed a scale to measure Mobile Phone Abuse, ATeMo, that takes into account the criteria for the diagnosis of substance use or addiction described in DSM-5 (American Psychiatric Association, 2002). The evaluation of craving was an important aspect of this questionnaire, as previously no measures existed that were consistent with the DSM-5 (American Psychiatric Association, 2012) criteria. The majority of measures had been developed based on the literature on substance use and addiction (Toda et al., 2004; Bianchi and Phillips, 2005; Rutland et al., 2007; Igarashi et al., 2008; Yen et al., 2009; Walsh et al., 2010; Chóliz, 2012; López-Fernández et al., 2012; Merlo et al., 2013; Chóliz et al., 2016), and the items in most of the previous instruments reflect the diagnostic criteria for substance use or addiction described in DSM-IV-TR (American Psychiatric Association, 2012). Based on the current findings we can conclude that the ATeMo questionnaire has satisfactory reliability and validity, having included craving as a diagnostic criteria for dependence.

AVAILABILITY OF DATA AND MATERIALS

R code and data are available from the authors under request.

ETHICS STATEMENT

This study was approved by the Research Ethics Committee from the Granada University. All procedures performed in our study involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

AUTHOR CONTRIBUTIONS

All the authors participated in the conception and design of the work, specifically MO-C and FL-T, conceived the original idea for the study, obtained funding and wrote the study protocol. MO-C manages the day to day running of the study, including all participant follow-up and IR-U and PH-T undertaked all data analyses. This study paper was written by FL-T, IR-U, and PH-T with input from all co-authors. All authors read and approved the final manuscript and believe that the manuscript represents valid work; carefully read and fully approve of it.

ACKNOWLEDGMENTS

REFERENCES

Abebe, D. S., Hafstad, G. S., Brunborg, G. S., Kumar, B. N., and Lien, L. (2015). Binge drinking, cannabis and tobacco use among ethnic Norwegian and ethnic minority adolescents in Oslo, Norway. *J. Immigr. Minor. Health* 17, 992–1001. doi: 10.1007/s10903-014-0077-9

American Psychiatric Association (2012). *DSM-5: The Future of Psychiatric Diagnosis*. Available inline at: http://www.dsm5.org/Pages/Default.aspx

American Psychiatric Association (2002). *Diagnostic and Statistical Manual of Mental Disorders IV-TR*. Washington, DC: American Psychiatric Association.

Bianchi, A., and Phillips, J. G. (2005). Psychological predictors of problem mobile phone use. *Cyberpsychol. Behav.* 8, 39–51. doi: 10.1089/cpb.2005.8.39

Billieux, J., Gay, P., Rochat, L., and Van der Linden, M. (2010). The role of urgency and its underlying psychological mechanisms in problematic behaviours. *Behav. Res. Ther.* 48, 1085–1096. doi: 10.1016/j.brat.2010.07.008

Billieux, J., Philippot, P., Schmid, C., Maurage, P., De Mol, J., and Van, D. L. (2014). Is dysfunctional use of the mobile phone a behavioural addiction? Confronting symptom based versus process based approaches. *Clin. Psychol. Psychother.* 22, 460–468. doi: 10.1002/cpp.1910

Billieux, J., Van Der Linden, M., and Rochat, L. (2008). The role of impulsivity in actual and problematic use of the mobile phone. *Appl. Cogn. Psychol.* 22, 1195–1210. doi: 10.1002/acp.1429

Biolcati, R. (2015). Women between normality and dependence: a study of problematic consumption and dependence in women aged 30-50 years. *J. Subst. Use* 20, 143–150. doi: 10.3109/14659891.2013.878759

Blasco, J., Martínez-Raga, J., Carrasco, E., and Didia-Attas, J. (2008). Atención y craving o ganas compulsivas. Avances en su conceptualización y su implicación en la prevención de recaídas. *Adicciones* 20, 365–376.

Chiu, S. I., Hong, F. Y., and Chiu, S. L. (2013). An analysis on the correlation and gender difference between college students' internet addiction and mobile phone addiction in Taiwan. *ISRN Addict.* 2013:360607. doi: 10.1155/2013/360607

Chóliz, M. (2012). Mobile-phone addiction in adolescence: the test of mobile phone dependence (TMD). *Prog. Health Sci.* 2, 33–44.

Chóliz, M., Pinto, L., Phansalkar, S. S., Corr, E., Mujjahid, A., Flores, C., et al. (2016). Development of a Brief Multicultural Version of the Test of Mobile Phone Dependence (TMD brief) Questionnaire. *Front. Psychol.* 7:650. doi: 10.3389/fpsyg.2016.00650

Contreras-Rodríguez, O., Albein-Urios, N., Vilar-López, R., Perales, J. C., Martínez-González, J. M., Fernández-Serrano, M. J., et al. (2016). Increased corticolimbic connectivity in cocaine dependence versus pathological gambling is associated with drug severity and emotion-related impulsivity. *Addict. Biol.* 21, 709–718. doi: 10.1111/adb.12242

De-Sola, J., Talledo, H., Rubio, G., and de Fonseca, F. R. (2017a). Psychological factors and alcohol use in problematic mobile phone use in the Spanish population. *Front. Psychiatry* 8:11. doi: 10.3389/fpsyt.2017.00011

De-Sola, J., Talledo, H., Rubio, G., and de Fonseca, F. R. (2017b). Development of a mobile phone addiction craving scale and its validation in a spanish adult population. *Front. Psychiatry* 8:90. doi: 10.3389/fpsyt.2017.00090

DeVellis, R. F. (2012). *Scale Development, Theory and Applications*. Los Angeles, CA: SAGE.

Fargues, M. B., Lusar, A. C., Jordania, C. G., and Sánchez, X. C. (2009). Validación de dos escalas breves para evaluar la adicción a internet y el abuso de móvil. *Psicothema* 21, 480–485.

Fishman, J. A., and Galguera, T. (2003). *Introduction to Test Construction in the Social and Behavioral Sciences*. Lanham: Rowman And Littlefield Publishers, Inc.

Foerster, M., Roser, K., Schoeni, A., and Röösli, M. (2015). Problematic mobile phone use in adolescents: derivation of a short scale MPPUS-10. *Int. J. Public* 60, 277–286. doi: 10.1007/s00038-015-0660-4

Furr, M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. Los Angeles, CA: SAGE Publications.

Gallimberti, L., Buja, A., Chindamo, S., Terraneo, A., Marini, E., Rabensteiner, A., et al. (2016). Problematic cell phone use for text messaging and substance abuse in early adolescence (11- to 13-year-olds). *Eur. J. Pediatr.* 175, 355–364. doi: 10.1007/s00431-015-2645-y

Grant, J. E., Schreiber, L. R., and Odlaug, B. L. (2013). Phenomenology and treatment of behavioural addictions. *Can. J. Psychiatry* 58, 252–259. doi: 10.1177/070674371305800502

Ha, J. H., Chin, B., Park, D. H, Ryu, S. H., and Yu, J. (2008). Characteristics of excessive cellular phone use in korean adolescents. *Cyber. Psychol. Behav.* 11, 783–784. doi: 10.1089/cpb.2008.0096

Hernández, A., Muñiz, J., and García, E. (2000). Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema* 12, 288–291.

Igarashi, T., Motoyoshy, T., Takai, J., and Yoshida, T. (2008). No mobile, no life: Self-perception and text-message dependency among Japanese high school students. *Comput. Hum. Behav.* 24, 2311–2324. doi: 10.1016/j.chb.2007.12.001

Jenaro, C., Flores, N., Gómez-Vela, M., González-Gil, F., and Caballo, C. (2007). Problematic internet and cell-phone use: psychological behavioral, and health correlates. *Addict. Res. Theory* 15, 309–320. doi: 10.1080/16066350701350247

Jiang, Z., and Shi, M. (2016). Prevalence and co-occurrence of compulsive buying, problematic internet and mobile phone use in college students in Yantai, China: relevance of self-traits. *BMC Public Health* 16:1211. doi: 10.1186/s12889-016-3884-1

Jöreskog, K. G., and Sörbom, D. (1996). *LISREL 8 User's reference guide*. Chicago, IL: Scientific Software.

Kawasaki, N., Tanei, S., Ogata, F., Burapadaja, S., Loetkham, C., Nakamura, T., et al. (2006). Survey on cellular phone usage on students in thailand. *J. Physiol. Anthropol.* 25, 377–382. doi: 10.2114/jpa2.25.377

Kwon, M., Kim, D. J., Cho, H., and Yang, S. (2013). The smartphone addiction scale: development and validation of a short version for adolescents. *PLoS ONE* 8:83558. doi: 10.1371/journal.pone.0083558

Larsen, C. V. L., Curtis, T., and Bjerregaard, P. (2013). Harmful alcohol use and frequent use of marijuana among lifetime problem gamblers and the prevalence of cross-addictive behaviour among green land inuit: Evidence from the cross-sectional inuit health in transition green land survey 2006-2010. *Int. J. Circumpolar Health* 72, 1–9. doi: 10.3402/ijch.v72i0.19551

Lee, K., Lee, H. K., Gyeong, H., Yu, B., Song, Y. M., and Kim, D. (2013). Reliability and validity of the Korean version of the Internet Addiction Test among college students. *J. Korean Med. Sci.* 28, 763–768. doi: 10.3346/jkms.2013.28.5.763

Leeman, R. F., and Potenza, M. N. (2013). A targeted review of the neurobiology and genetics of behavioural addictions: an emerging area of research. *Can. J. Psychiatry* 58, 260–273. doi: 10.11772F070674371305800503

Leung, L. (2008). Linking psychological attributes to addiction and improper use of the mobile phone among adolescents in Hong Kong. *J. Child Media* 2, 93–113. doi: 10.1080/17482790802078565

Lin, T. T. C., Chiang, Y., and Jiang, Q. (2015). Sociable people beware? Investigating smartphone versus no smartphone dependency symptoms among young Singaporeans. *Soc. Behav. Personal.* 43, 1209–1216. doi: 10.2224/sbp.2015.43.7.1209

López-Fernández, O., Honrubia-Serrano, M., and Freixa-Blanxart, M. (2012). Adaptación española del "Mobile Phone Problem Use Scale" para población adolescente. *Adicciones* 24, 123–130. doi: 10.20882/adicciones.104

López-Torrecillas, F., García, J, Godoy, J. F., García, M. P., Izquierdo, D. G., and Sánchez-Barrera, M. B. (2000). Variables modulating stress and coping that discriminate drug consumers from low or nondrug consumers. *Addict. Behav.* 25, 161–165. doi: 10.1016/S0306-4603(98)00121-X

Marlatt, G. A., Baer, J. S., Donovan, D. M., and Kivlahan, D. R. (1988). Addictive behaviors: etiology and treatment. *Annu. Rev. Psychol.* 39, 223–252.

Mattebo, M., Tydén, T., Häggström-Nordin, E., Nilsson, K. W., and Larsson, M. (2013). Pornography consumption, sexual experiences, lifestyles, and self-rated health among male adolescents in Sweden. *J. Dev. Behav. Pediatr.* 34, 460–468. doi: 10.1097/DBP.0b013e31829c44a2

Mentzoni, R. A., Brunborg, G. S., Molde, H., Myrseth, H., Skouverøe, K. J. M., Hetland, J., et al. (2011). Problematic video game use: estimated prevalence and associations with mental and physical health. *Cyberpsychol. Behav. Soc. Netw.* 14, 591–596. doi: 10.1089/cyber.2010.0260

Merlo, L. J., Stone, A. M., and Bibbey, A. (2013). Measuring problematic mobile phone use: Development and preliminary psychometric properties of the PUMP scale. *J Addict.* 2013:912807. doi: 10.1155/2013/912807

Míguez, C. M., and Becoña, E. (2015). Do cigarette smoking and alcohol consumption associate with cannabis use problem gambling among Spanish adolescents? *Adicciones* 27, 8–16.

Morata-Ramírez, M., Holgado-Tello, F., P., Barbero-García, I., and Mendez, G. (2015). Confirmatory factor analysis. Recommendations for unweighted least squares method related to Chi-Square and RMSEA. *Accion Psicol.* 12, 79–90. doi: 10.5944/ap.12.1.14362

Mudry, T. E., Hodgins, D. C., el-Guebaly, N., Cameron Wild, T., Colman, I. B., Schopflocher, D. et al. (2011). Conceptualizing excessive behaviour syndromes: a systematic review. *Curr. Psychiatry Rev.* 7, 138–151. doi: 10.2174/157340011796391201

Navas, J. F., Billieux, J., Perandrés-Gómez, A., López-Torrecillas, F., Cándido, A., and Perales, J. C. (2017). Impulsivity traits and gambling cognitions associated with gambling preferences and clinical status. *Int. Gambl. Stud.* 17, 102–124. doi: 10.1080/14459795.2016.1275739

Osuch, E., Vingilis, E., Ross, E., Forster, C., and Summerhurst, C. (2013). Cannabis use addiction risk and functional impairment in youth seeking treatment for primary mood or anxiety concerns. *Int. J. Adolesc. Med. Health* 25, 309–314. doi: 10.1515/ijamh-2013-0067

Pedrero-Pérez, E. J., Rodríguez-Monje, M. T., Gallardo- Alonso, F., Fernández- Girón, M., Pérez- López, M., and Chicharro- Romero, J. (2007). Validación de un instrumento para la detección de trastornos de control de impulsos y adicciones: el MULTICAGE CAD-4. *TrastorAdic* 9, 269–278. doi: 10.1016/S1575-0973(07)75656-8

Roberts, J. A., Petnji-Yaya, L. H., and Manolis, C. (2014). The invisible addiction: Cell-phone activities and addiction among male and female college students. *J. Behav. Addict.* 3, 254–265. doi: 10.1556/JBA.3.2014.015

Roser, K., Schoeni, A., Foerster, M., and Röösli, M. (2016). Problematic mobile phone use of Swiss adolescents: is it linked with mental health or behaviour? *Int. J. Public Health* 61, 307–315. doi: 10.1007/s00038-015-0751-2

Rosnow, R. L., and Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: general procedures for research consumers. *Psychol. Methods* 1, 331–340. doi: 10.1037/1082-989X.1.4.331

Ruiz-Olivares, R., Lucena, V., Pino, M. J., and Herruzo, J. (2010). Análisis de comportamientos relacionados con el uso/abuso de internet, teléfono móvil, compras y juego en estudiantes universitarios. *Adicciones* 22, 301–310. doi: 10.20882/adicciones.171

Rutland, J. B., Sheets, T., and Young, T. (2007). Development of a scale to measure problem use of short message service: the SMS problem use diagnostic questionnaire. *Prog. Health Sci.* 10, 841–843. doi: 10.1089/cpb. 2007.9943

Ryu, Y. S., and Smith-Jackson, T. L. (2006). Reliability and Validity of the Mobile Phone Usability Questionnaire (MPUQ). *J. Usab. Stud.* 2, 39–53.

Schepers, J. (2009). *Test Construction: Theory and practice. A Course for Beginners.* Saarbrucken: VDM, Verlag Dr. Muller.

Schuster, R. M., Mermelstein, R., and Wakschlag, L. (2013). Gender-specific relationships between depressive symptoms, marijuana use, parental communication and risky sexual behavior in adolescence. *J. Youth Adolesc.* 42, 1194–1209. doi: 10.1007/s10964-012-9809-0

Takao, M., Takahashi, S., and Kitamura, M. (2009). Addictive personality and problematic mobile phone use. *Cyberpsychol. Behav.*, 2, 501–507. doi: 10.1089cpb.2009.0022

The International Telecommunication Union (2016). Available online at: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures'2016.pdf

Toda, M., Monden, K., Kubo, K., and Morimoto, K. (2004). Cellular phone dependence tendency of female university students. *JPN J. Hyg.* 59, 383–386. doi: 10.1265/jjh.59.383

Van Rooij, A. J., Kuss, D. J., Griffiths, M. D., Shorter, G. W., Schoenmakers, M. T., and Mheen, V. D. (2014). The (co-)occurrence of problematic video gaming, substance use, and psychosocial problems in adolescents. *J. Behav. Addict.* 3, 157–165. doi: 10.1556/JBA.3.2014.013

Vanyukov, M. M., Tarter, R. E., Kirillova, G. P., Kirisci, L., Reynolds, M. D., Kreek, M. J., et al. (2012). Common liability to addiction and "gateway hypothesis": theoretical, empirical and evolutionary perspective. *Drug Alcohol Depend.* 123, 3–17. doi: 10.1016/j.drugalcdep.2011.12.018

Verdejo-García, A., López-Torrecillas, F., Carmen, O. G., and Pérez-García, M. (2004). Clinical implications and methodological challenges in the study of the neuropsychological correlates of cannabis, stimulant, and opioid abuse. *Neuropsychol. Rev.* 14, 1–41. doi: 10.1023/B:NERV.0000026647.71528.83

Viola, T. W., Tractenberg, S. G., Wearick-Silva, L., de, O. R., Pezzi, J., and Grassi-Oliveira, R. (2014). Long-term cannabis abuse and early-onset cannabis use increase the severity of cocaine withdrawal during detoxification and rehospitalization rates due to cocaine dependence. *Drug Alcohol Depend.* 144, 153–159. doi: 10.1016/j.drugalcdep.2014.09.003

Vorspan, F., Mehtelli, W., Dupuy, G., Bloch, V., and Lépine, J. (2015). Anxiety and substance use disorders: co-occurrence and clinical issues. *Curr. Psychiatry Rep.* 17:4. doi: 10.1007/s11920-014-0544-y

Walsh, S. P., White, K. M., McD., and Young, R. (2010). Needing to connect: the effect of self and others on young people's involvement with their mobile phones. *Aust. J. Psychol.* 62, 194–203. doi: 10.1080/00049530903567229

Walther, B., Morgenstern, M., and Hanewinkel, R. (2012). Co-occurrence of addictive behaviours: personality factors related to substance use, gambling and computer gaming. *Eur. Addict. Res.* 18, 167–174. doi: 10.1159/000335662

Weinstein, A., and Lejoyeux, M. (2015). New developments on the neurobiological and pharmaco-genetic mechanisms underlying internet and videogame addiction. *Am. J. Addict.* 24, 117–125. doi: 10.1111/ajad.12110

Xu, H., Wu, X., Lan, Y., and Chen, Y. (2008). Development of mobile phone dependence inventory for college students. *Chin. J. Clin. Psychol.* 16, 26–27.

Yang-Wallentin, F., Jöreskog, K. G., and Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Struct. Equ. Model.* 17, 392–423. doi: 10.1080/10705511.2010.489003

Yau, Y. H., Crowley, M. J., Mayes, L. C., and Potenza, M. N. (2012). Are Internet use and videogame-playing addictive behaviors? Biological, clinical and public health implications for youths and adults. *Min. Psichiatr.* 53, 153–170.

Yen, C. F., Tang, T. J., Yen, J. Y., Lin, H. C., Huang, C. F., Liu, S. C., et al. (2009). Symptoms of problematic cellular phone use, functional impairment and its association with depression among adolescents in southern Taiwan. *?J. Adolesc.* 32, 863–873. doi: 10.1016/j.adolescence.2008.10.006

Peace Data Standard: A Practical and Theoretical Framework for Using Technology to Examine Intergroup Interactions

Rosanna E. Guadagno[1]*, Mark Nelson[1] and Laurence Lock Lee[2]

[1] Peace Innovation Lab at Stanford, Stanford University, Stanford, CA, United States, [2] SWOOP Analytics, Sydney, NSW, Australia

The current paper presents a theoretical framework for standardizing Peace Data as a means of understanding the conditions under which people's technology use results in positive engagement and peace. Thus, the main point of our paper is that Big Data can be conceptualized in terms of its value to peace. We define peace as a set of positive, prosocial *behaviors* that maximize mutually beneficial positive outcomes resulting from interactions with others. To accomplish this goal, we present hypothetical and real-world, data driven examples that illustrate our thinking in this domain and present guidelines for how to identify, collect, utilize, and evaluate Peace Data generated during mediated interactions and further suggest that Peace Data has four primary components: group identity information, behavior data, longitudinal data, and metadata. This paper concludes with a call for participation in a Peace Data association and suggested for guidelines for how scholars and practitioners can identify Peace Data in their own domains. Ethical considerations and suggestions for future research are also discussed.

Keywords: peace, Big Data, social media, research methods, data science

"People are people so why should it be, You and I should get along so awfully?"

–*Depeche Mode* (Gore, 1984)

INTRODUCTION

Even the most casual perusal of the news headlines confirms what the classic 1980s new wave band *Depeche Mode* put so eloquently; we humans often focus on our differences, from largely visible social categories such as age, gender, race, ethnicity, language, and socio-economic status, to less visible characteristics such as sexual orientation, disability status, occupation, education level, religious affiliation, and political orientation (Brewer, 1991). Similarly, research findings in the psychological sciences support this notion. For instance, Henri Tajfel established that by randomly sorting people into meaningless groups (i.e., the "blue" vs. "green" groups commonly referred to as the minimal group paradigm), people begin to prefer their group over the outgroup. Thus,

dividing the world into "us" vs. "them" is as automatic a process as blinking our eyes and has been shown to broadly affect people's perception of others (Tajfel, 1970; Balliet et al., 2014). People automatically categorize others by visible social categories (i.e., gender and ethnicity), as part of this process (Brewer, 1988). On the Internet, these psychological processes are magnified by the one-to-many methods of communication typified by Web 2.0 technologies such as social media and review aggregating websites (Amichai-Hamburger, 2005, 2015).

While much research has demonstrated the many ways these group differences lead to conflict (e.g., Allport, 1954/1979; Sherif, 1961; Hogg, 2016), this paper takes a different approach. We suggest that the precise group differences that otherwise would cause conflict, can instead generate prosocial behavior and new wealth through structured *engagement episodes* – interactions mediated by various networked technologies (e.g., crowdsourcing applications, social media, texting, email). This is especially the case with services that facilitate person-to-person financial interactions, such as when people rent each other's cars and homes, thereby creating new wealth, new relationships, and new opportunities to form friendships. We suggest that these services yield mutual benefit in excess of the cost of engagement for both interactants. It is these discrete episodes of engagement, and the mutually beneficial interactions they comprise that, we argue, constitute meaningful positive peace that can be measured in useful ways.

OVERVIEW OF THE PRESENT PAPER

How do we observe and measure the amount of positive engagement generated by different actors and applications? We suggest that big data can be conceptualized in terms of its ability to understand and initiate peaceful interactions. Thus, our theoretical framework for Peace Data provides a starting point. This initial version of the Peace Data Standard generalizes, and applies to ANY mediating technology, from old-fashioned "landline" phone calls, or Internet-based financial transactions, to the latest connection anyone makes on crowdsourced social media or dating applications such as Snapchat or Tinder. In the present paper, we focus on these questions by presenting our theoretical perspective on Peace Data, presenting guidelines for the identification, collection, examples of, uses for, and value of Peace Data generated during mediated interactions. These guidelines make hypotheses about the four primary facets of Peace Data: group identity information, behavior data, longitudinal data, and metadata. We further present several examples, hypothetical and genuine, of contexts in which data consistent with our proposed Peace Data Standard could be collected. Next, we conclude with a call for participation in a Peace Data association, suggesting guidelines for how both scholars and practitioners can identify Peace Data in their own organizations and datasets. Finally, we present suggested directions for future research, and a conclude with a preliminary discussion of the many ethical considerations in the collection and use of Peace Data.

Mediating technologies[1], which we define as technology that "acts as an intervening agent, augmenting our ability to engage positively with others (Quihuis et al., 2015)" take on the role of a social actor (e.g., Reeves and Nass, 1996), connecting individuals acting independently by supported coordinated behavior. Thus, mediating technologies are those that connect people – oftentimes strangers – from divergent backgrounds to facilitate positive engagement. We define engagement in terms of both the quantity and quality of interaction. It can be either positive, reflecting high quality and frequent interaction or negative, defined as low quality and low frequency of interaction (see **Figure 1** for a visual of this process). Thus, these mediating technologies enable people to rapidly discover, refine, scale, and simultaneously assess in real-time the quantity and quality of mutually beneficial interactions between any two groups or entities. This in turn allows PeaceTech entrepreneurs, scholars, and designers to rapidly design, test, and validate interventions that effectively transform these group differences into raw material for sustainable peace (in which mutual benefit is equal to the cost of engagement) and eventually scalable positive peace (in which mutual benefit exceeds the cost of engagement). Prior to introduction of contemporary mediating technologies, the vast majority of human interaction was not easily recorded. In today's world, we can easily record, analyze, and draw inferences – even in real-time – large samples of human interactions that occur via mediated-tech such as social media and mobile applications. We further suggest that these technological advances may even provide a means to remove resources and incentives from violent conflict situations, something we recommend as a direction for future research to explore.

---

[1]Note that we intentionally using the phrase "mediating technology" rather than "mediated technology" to suggest that, in addition to technology being used as a means of communication (as in the case of "mediated communication"), it can also serve as a virtual stand in for a mediator – an objective third party that assists with conflict resolution. That said, the phrase "mediated technology" also appears in the manuscript and it refers to the more traditional definition.

FIGURE 1 | The primary dimensions of peace data.
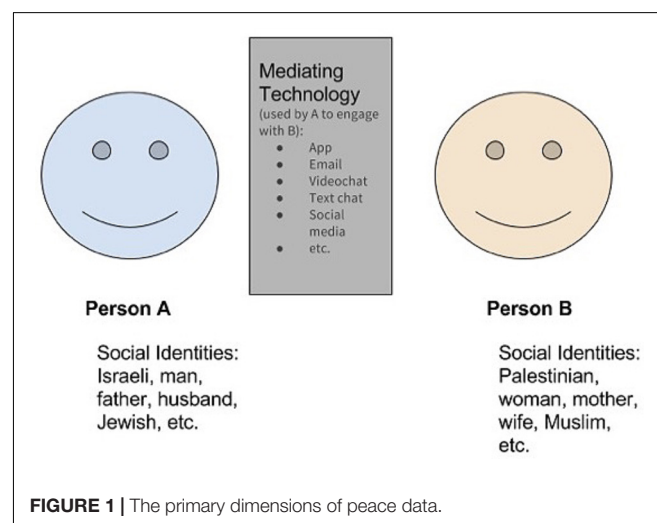
Peace Defined

Webster's dictionary defines peace as: "a pact or agreement to end hostilities between those who have been at war or in a state of enmity" (Peace, n.d.). Cohrs et al. (2013) define peace as: "not only the absence or minimization of violence but also the presence or development of harmonious relationships (Anderson, 2004) and social justice (Galtung, 1969)" (p. 590). Furthermore, other research differentiates between negative peace – the more traditional perspective that pertains to the reduction, cessation, and prevention of violence – and positive peace – relief from violence and the introduction of social justice (e.g., Christie and Montiel, 2013). Building on these definitions, in our work at the Peace Innovation Lab at Stanford, we take both a behavioral and positive perspective, defining peace as a suite of positive, prosocial *behaviors* that maximize mutually beneficial positive outcomes from interactions with others.

THE CONTACT HYPOTHESIS

In the present paper, we suggest that mediated-technology can be used to facilitate and measure peace, and specifically positive peace (that is, pro-social, and even mutually beneficial behavior across group boundaries). Given that it is widely established in psychology that, under the right conditions, contact between different groups can reduce intergroup conflict and facilitate positive interactions across group boundaries (Allport, 1954/1979), we suggest that if designed correctly, mediated-technology can increase positive peace. In support of this, research has shown that that mediated-contact, when properly designed and implemented can reduce intergroup conflict (Amichai-Hamburger et al., 2015; White et al., 2015). Specifically, both sets of the authors reviewed the results of several studies that illustrated the benefit of e-contact as an initial means of intergroup contact, particularly with respect to reducing intergroup bias and anxiety and increasing knowledge between groups. The authors theorized this is particularly effective when: e-contact takes place more than once at different time points, the interactants acknowledge both group similarities and differences, and the form of e-contact includes Allport's (1954/1979) conditions for reducing intergroup conflict: equal status, common goals, cooperation, and support from authority). A Meta-analysis of over 500 studies on the contact hypothesis has verified its effectiveness of contact theory in its ability to reduce prejudice (Pettigrew and Tropp, 2006).

One theme apparent in both reviews is the notion that medicated-intergroup contact will be most successful when the interaction is structured in nature (Amichai-Hamburger et al., 2015; White et al., 2015). For instance, Amichai-Hamburger et al. (2015) defines structured contact as intergroup contact in which group members are selected for participation, the numbers are equal for each group, the contact is observed. Amichai-Hamburger et al. (2015) further suggest that mediated intergroup contact can be effective because of the following seven characteristics: "anonymity, control over the physical exposure, control over the interaction, ease of finding similar

others, universal and constant availability and accessibility of the Internet, equality, and fun" (p. 517). Illustrative of this, the authors present examples of contemporary technology solutions built with the principles of contact theory. For instance, Games for Peace[2] recruits Israeli and Palestinian children to interact in various virtual environments designed with the principles of contact theory to counteract negative stereotypes that people often hold of members of other groups and facilitate peaceful relations. Similarly, the authors also review a program called The Peace Factory[3] which uses similar principles to foster peace in the Middle East by facilitating social media friendships between people from Middle Eastern countries in conflict. For instance, the Peace Factory launched a Facebook group called "Israel loves Iran"[4] that provides a safe and public space for people from these two cultures to connect, communicate, and form friendships.

Our own lab also previously partnered with Facebook to create a page devoted to emphasizing the social media friendships created across conflict groups. Facebook's Peace page (called Peace Dot[5]) reports the number of friend requests accepted between conflicting groups in real-time. It updates every 24 h and currently displays data on the friendships created across the following groups: Israel vs. Palestine, Pakistan vs. India, and Ukraine vs. Russia. This page also emphasizes the point that, even when conflict between these groups are high, people from these groups are forming more friendships than they are harming each other. For instance, Quihuis et al. (2015) reported that during a 2012 resurgence of Israeli-Palestinian violence, there were still over 13 times as many friendships formed on Facebook for each reported injury or death (see **Figure 2** for a screenshot of Peace Dot).

Other research has similarly found that technology can serve as an effective means of producing peaceful relations and reducing prejudice. For instance, Walther et al. (2015) placed Israeli and Palestinian students in a year-long intervention designed with the principles of contact theory. To examine this, groups of six students were placed in computer-based discussion groups as part of their participation in a course on Advanced Educational Environments. These students were members of three different in groups: secular Jewish, religious Jewish, and Arab (Muslim). The authors also recruited a control group that did not participate in the intervention. All participants filled out a series of pre- and post-measures of prejudice toward members of all three groups. Their results revealed that after communicating in their small groups throughout the entire academic year, participants were significantly less prejudiced toward the outgroups relative to both their pre-test scores and compared to the control group who did not participate in the intervention. Similarly, Cao and Lin (2017) reported that visual anonymity during interactions between people from different groups was effective in decreasing prejudice toward a specific outgroup member but was not effective in improving intergroup relations more broadly. However, when the authors added a

---

[2]http://gamesforpeace.org
[3]http://thepeacefactory.org
[4]https://www.facebook.com/israellovesiran/
[5]https://www.facebook.com/peace
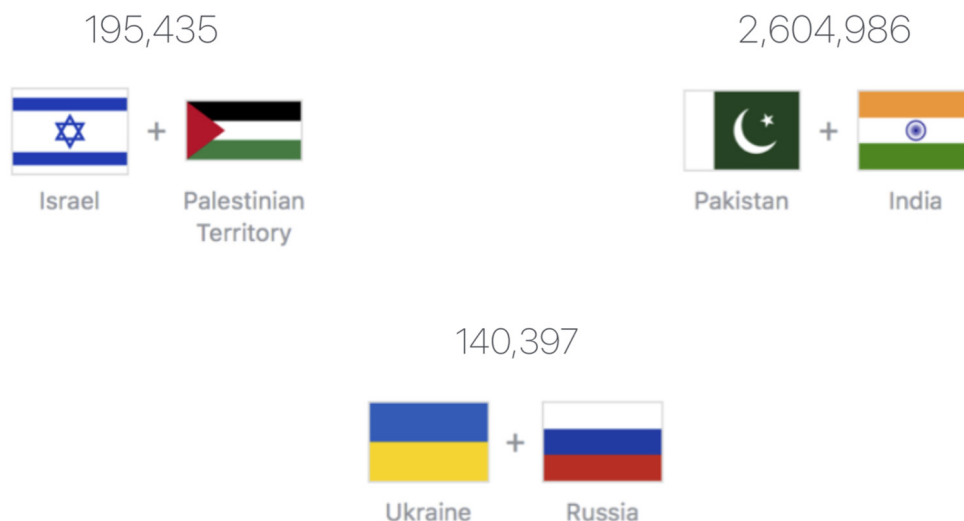
FIGURE 2 | Screenshot of Peace Dot taken on February 19, 2018.

video-based chat, they found that contact did improve attitudes toward different groups.

Taken together, the extant literature demonstrates the effectiveness of Internet communications technology designed using the principles of the contact hypothesis to facilitate peaceful interactions and reducing prejudice. However, this research focuses solely on the conditions under which communications technology can be used to improve intergroup relations but

does not does not examine how technology can serve as a mediator of peace for face-to-face interactions. We suggest that in the today's app-centric Internet, this is a more realistic use of technology in bridging the divide between groups. Indeed, Amichai-Hamburger and McKenna (2006) suggest that the use of technology as a mediator for conflict resolution occurs on a continuum that, if implemented correctly, can eventually transition into peaceful face-to-face interactions. Unfortunately, to date, little research has studied this transition. Given the paucity of research that studies the role of technology as a mediator of peaceful interactions in person, we provide hypothetical examples of the role mediating technologies play in facilitating positive, peaceful social interactions appear in **Table 1**. As both examples illustrate, ridesharing and the crowdsourcing of short- and long-term lodging by popular applications such as Airbnb, VRBO, Roomorama, HomeAway, etc. are both revolutionizing the way people travel and are also examples of the many mediating technologies that provide opportunities for positive engagement. Similarly, this paper was written collaboratively using a popular shared word processing software, and the widely used crowdsourcing platform Amazon Mechanical Turk provides similar positive engagement and new wealth between scientists in need of human research participants (e.g., Buhrmester et al., 2011) and between artists who crowdsource to design their art installations (burrough, 2016).

FOUR DIMENSIONS OF PEACE DATA

As we conceptualize it, standardizing methods for the collection of Peace Data will enhance the ability of researchers, innovators, and practitioners to engage in an open and transparent examination of the effectiveness of mediating technology to promote peace and related outcomes. Below, we further expand on four psychological and methodological factor that should be considered as standard practices in Peace Data emerge. These topics are: group identity, the behavioral nature of Peace Data, the potential for longitudinal data collection, and the use of metadata as part of the Peace Data Standard. We further suggest that, while having all four dimensions represented in a dataset, having data which contains any one of these dimensions can be valuable in modeling the role of mediated communication in facilitating peace.

Group Identity
Group identity data refers to the social categories people associate with themselves (e.g., "I am a student, a feminist, a parent," Brewer, 1991), the groups that people sort others into both in terms of ingroup (e.g., "is that other person a member of my group?"; Tajfel and Turner, 1986) and in terms of the social categories that are easily observable (e.g., sex, ethnic background, age; Brewer, 1991). We hypothesize that new structural and statistical identities may also now be discoverable, through the analysis of big data using computer algorithms or machine learning. We further hypothesize that these identities may not discoverable by human observation

and may instead emerge from consistency in one's behavior. Illustrative of this, research indicates that certain technological solutions can: extrapolate people's personality characteristics (i.e., extraversion) from the content and frequency of people's online posts (e.g., Azucar et al., 2018), identify easily to persuade people from their online shopping behavior (Kooti et al., 2016), and develop profiles to differentiate customers into different groups based on their purchasing trends (Wen et al., 2018). People's varying group identities are akin to *difference boundaries* – group difference categories that reflect a single or many, possibly nested differences in social identity (Leigh Star, 2010).

Note that, as the concealable stigma literature illustrates, not all social identities need to be known by both parties to affect a social interaction. Stigma refers to a social identity that is valued lower than other group identities (Crocker et al., 1998). Concealable stigmas are such undervalued social identities that can be hidden (Goffman, 2009). Identifying stigmatized social identities is somewhat context dependent and often results in discrimination against members of this group. Membership in a stigmatized group has been shown to negatively affect people's physical and mental health (Major and O'Brien, 2005). Thus, people who are members of a stigmatized group (e.g., a chronic illness that is not readily observable) that can be concealed, may choose to do so to avoid these negative outcomes. Nonetheless, people with concealable stigmas carry the knowledge of their group membership with this knowledge can influence the expectations they carry into interactions, their responses to their interaction partners, and affect the impressions people for of them (Goffman, 2009). Related to the present paper, MacInnis and Hodson (2015) assert that mediated communication may provide a safe space for cross group relationships to forms. In their study, these examined how revealing a concealable stigma via mediated communication affected relationship formation between two interactants, one with a concealable stigma and one without. Their results indicated that revealing a stigmatized social identity early on in a relationship facilitates the formation of a stronger cross group relationship. Related to Peace Data, we hypothesize that there may be other, unknown or yet to be discovered social identities that may be revealed through the use of Big Data and Machine Learning to techniques to further understand how membership such a group affects the likelihood of an engagement episode resulting in an increase in peace for the interacts.

Behavior Data
The second dimension of Peace Data is that it reflects people's actual behavior. As much research indicates, directly recording people's actions through *behavioral measures*, often provides unique insights about human social interaction (e.g., Baumeister et al., 2007; Lewandowski and Strohmetz, 2009). Prior to the era of big data driven by people's increasing reliance on interactive, smart technologies, collecting meaningful samples of people's behavior was considered too time consuming or labor intensive for many researchers to consider employing it in their research. Because increasingly ubiquitous sensors in our environment, with increasingly nuanced resolution and sensitivity, can now

TABLE 1 | Hypothetical examples of peace data.

Example 1. Ridesharing Applications	Example 2. Crowdsourced Lodging
On a recent trip to the airport, one of us (MN), used a popular ridesharing application on his mobile device to arrange his ride. Ride sharing applications (e.g., Gett, Uber, Juno, Lyft) have been gaining in popularity over traditional taxis, yet are not yet a widely accepted replacement for taxis (e.g., research reports that only 15% of American adults have used a ridesharing service, Smith, 2016). Despite this, growing anecdotal and scholarly evidence as well as an increasing market share all indicate that people prefer ridesharing to taxis (Deloitte Access Economics, 2016; Smith, 2016). In MN's case, he spent most of the 42-min ride getting to know his driver, a young woman we shall call Ayanna[1]. Ayanna was a 19-year-old Muslim woman from Somalia who proudly wore a headscarf. As an older white man from Canada without a religious preference, MN seemingly had little in common with Ayanna. The two would not likely have ever encountered each other if not for the ridesharing application (also referred to as an App). As the ride progressed, MN asked whether Ayanna's family approved of her job. Ayanna replied that her father and uncle were both taxi drivers out of a local airport and would never dream of letting her drive a taxi, not least because of the danger to her. However, with the ridesharing app tracking Ayanna's location, her passengers' identities and their driver-provided ratings, their pick up location, their intended destination, and the actual drop off location, the increased safety afforded by the technology made her family more than comfortable–they also thought this job would provide a great opportunity for her to help Americans personally experience how a black Muslim woman refugee from Somalia can be a valued and contributing member of society. So, for the cost of a $62 ride to the airport, the app enabled MN and Ayanna to discover each other, communicate both need and intent to meet that need, coordinate activity, trust each other even though they had never met, complete a mutually beneficial transaction, settle the transaction, and monetize the benefit. In the process, they created *and distributed* new economic, social, and arguably psychological wealth that could not have been generated without the mediating app. But unlike any other time in human history, all of these technology-enabled benefits are being passively measured and recorded–*in real-time*. The result? Over the course of 42 min, the ridesharing app enabled them to generate and record measurable positive engagement across ethnic, racial, religious, gender, nationality, language, and age boundaries–and measure some of the first-order economic and social impact of that engagement. Thus, this technology facilitated intergroup contact that was mutually beneficial for both parties, which as empirical evidence demonstrates, lead to a reduction in intergroup conflict (McKeown and Dixon, 2017).	When Pero and Gemma first learned about crowdsourced lodging, they decided to rent out part of their home as a way to supplement their income. Initially, that is exactly what happened. Not only did the couple increase their incomes, by sharing their home in this manner, they met people from all over the world, adding many of their guests to their circle of friends. Over time, as crowdsourced lodging became more popular through the use of various lodging-based social media sites (airbnb, vrbo, etc.), more and more of their neighbors started renting out all or part of their homes through these sites. Initially this was a boon to the local economy as more tourists came to visit the sights their European city had to offer. However, over time, Pero and Gemma started realizing that there were unintended consequences to their decision for their overall community. Local businesses run by their neighbors were among the first casualties of the share economy. First the local hardware store was replaced by a chain store that rents bicycles to tourists. This was followed by a number of local businesses being replaced by other tourist-centered (and expensive!) stores, restaurants, and services. Families who had lived in Pero and Gemma's town for generations soon found that they could not afford housing in their city and many ended up moving to a nearby town without anything to draw in tourists. Thus, what started as a way to make some extra money and meet new people ended up disrupting the economy of local community and disrupted the bonds within the community as well. These unforeseen negative consequences of the share economy have led some to argue that crowdsourced lodging should be beneficial to the overall community not just the people renting out their property (van der Zee, 2016, October 6). Imagine Elia encounters Andres while seeking to rent his spare room for a weekend, through a home-sharing application. Andres' posting of his spare room for rent is the initial *episode* of engagement between them, even though he has not met Elia yet and does not know who will reply, his willingness to engage is an important behavioral signal about both characteristics of his salient group identities, and about potential for engagement. Then Elia's initial message to Andres about renting his room, gives us our first data point about the actual *relationship* not only between them, but also potentially between each of the groups Elia and Andres are members of. This includes both their broad, obvious group identities (e.g., men vs. women) and their more nuanced and previously much less visible group identities (e.g., Andres is a member of: Ph.Ds., retirees, atheists, Columbians, fathers of one, vs. Elia's corresponding group identities of: 8th grade educations, working, Christians, Syrians, mothers of six). When Andres replies, perhaps about the dates the room is available, we can observe a second engagement episode, and these two episodes, one in each direction, create an *interaction*. Note that this is again an interaction both between them as individuals, *and* the groups they belong to (See **Figure 1**). Now, as we see a series of interactions between them over time, as they perhaps discuss price, amenities, and the duration of Elia's stay, we can begin to quantify and model some precise qualities of their *relationship* that have never been visible before. Next, if we want to know more about the state of affairs between any two of the groups Elia and Andres are each a member of, we can aggregate all the relationship data, from other group members like them, to say something empirical about the *group dynamics* between, groups such as: Christians and Atheists, or Syrians and Colombians. Finally, depending on the interaction context, organizations may be able to attach an economic value to the interaction. In some cases, corporations may know this information, in others they may not; and when it is known, this value may change from before to after the interaction.

[1]Note her real name but instead selected in homage to one of first author's childhood friends.

passively, unobtrusively, and automatically detect many kinds of human behavior, and especially human *social* behavior, the use of big data techniques to collect and analyze people's behavior has led to new and novel insights about people. For instance, one study found that data on the frequency of different Internet search terms entered into Google were predictive of subsequent behavioral trends (Choi and Varian, 2012). Specifically, the

researchers reported that this method could be used to predict seasonal variability in behaviors such as: visiting Hong Kong on vacation, unemployment trends (from searches for information on filing for unemployment insurance) and auto purchase trends (from search terms related to different types of cars). Applied to the question of Peace Data, we hypothesize that measuring actual behavior is essential for our understanding of the processes

involved in producing and predicting peaceful interactions between people in from different groups.

Longitudinal Data

A tertiary dimension of Peace Data is that it can be longitudinal in nature. Thus, scholars can use our model of peace to examine how prolonged positive engagement over time can facilitate peaceful outcomes between interactants. While this can be based on as little as a single engagement episode (e.g. clicking "like") or social interaction (e.g., a Lyft ride) between two people, it can scale to multiple exchanges between the same people over time (e.g., renting the same vacation rental each year through the same crowdsourced lodging application) or can be applied to different people and different contexts over time. No matter how it is applied, we hypothesize that the longitudinal aspect of Peace Data presents an opportunity for people to understand how long-term positive engagement can facilitate peace.

Related to this is the question of how episodes of engagement between a pair of people affects others in their same situation or social system? Dynamical systems approaches to modeling human social behavior (e.g., Nowak et al., 2013) suggest that, over time, peaceful outcomes through positive engagement should spread throughout a social system and that it will spread fastest to the people most closely associated with the initial pair of interactants (e.g., Okdie et al., 2018, Unpublished).

Metadata

The final aspect of Peace Data is that it can also include metadata, defined as data that provides information on other data (Ball, 2017). This can include aggregate data and/or descriptive statistics that provide group-level information on the mechanisms involved in achieving peaceful outcomes from social interactions mediated by technology. Metadata such as this allows scholars to track interactions across time and compare outcomes to other contexts, people, groups, and interactions. For instance, assessing the latency in messages sent between people through an app or social media site may reveal the extent to which people favor their own group members in a given setting of context. Currently the extant research on metadata in psychological science is scant. However, we hypothesize that, depending on the scope of the Peace Data collected, metadata can provide information about peace through positive engagement on different scales of measurement, for instance, at the level of an engagement episode or longer interaction and at the level of an individual pair or a larger group.

IDENTIFYING PEACE DATA

How can scholars, innovators, and practitioners identify Peace Data? We suggest that any data that meets the criteria established in the four dimensions of Peace Data described is peace data. Consider this hypothetical interaction between two hypothetical people: person A and person B. Person A (Elsa, for the purposes of a generic example) has a variety of shared and unique group

identities. For instance, she could be an African, Christian, mother of six. Person B (Toby, for the purposes of a generic example) may be a retired, atheist, Latino grandfather with a Ph.D. They may be from different countries, sexes, ethnicities, religious groups, and/or education levels (what we call *difference boundaries*; Leigh Star, 2010), but may then connect through some mediating technology. Difference boundaries can vary in many ways. For instance, a broad distinction between people is their gender[6] (men vs. women). This distinction glosses over other group differences that may affect the outcome of the engagement episode. As the running example illustrates Elsa and Toby not only differ in gender, they also differ in religion, nationality, education, family size, and likely age. This raises concerns with respect to the generalizability of research findings to other people and engagement episodes and calls for guidelines for the ethical design of technologies that automatically detect and assign group identities to people.

PEACE DATA TYPES

What kind of information do we recommend scholars, practitioners, and innovators collect in conjunction with Peace Data? To fully address this question, we must first review the relationship between people's group identities, and how inter-group or cross-group-boundary engagement shapes their identities and affects their behavior. As indicated above, people have a tendency to divide the world into two groups: people who are part of their group, referred to as their "ingroup" vs. people who are not, referred to as an "outgroup" (Tajfel and Turner, 1986). The dividing line is typically based on salience – the social identity that is either most visible or most relevant to the context (Onorato and Turner, 2004). For instance, in a large group of men, two women with nothing else in common will likely prefer each other to other group members as gender is the salient social identity for that context. These same two women may feel anything but camaraderie when their nationality or political affiliations are the social identities more central to the social context. Research has shown that this is the case, particularly when interaction is mediated by technology (e.g., Spears et al., 2002).

Table 2 shows examples of how Peace Data could be collected and formatted for data analysis. As sensor technology advances, we hypothesize that more precise detection, measurement, and modeling of prosocial, peaceful behavior in mediated interpersonal interactions will become possible. If enough people opt in to our proposed Peace Data Standard, we could follow a particular chain of engagement episodes between the same people, across platforms (e.g., eBay and PayPal). In this case the episodes are each visible on the respective platform, but the interaction is only visible across platforms.

Additionally, ascertaining the level of analysis is also important to the collection of Peace Data. Since utility, value, and richness of the data likely increases longitudinally, we

---

[6]This example pertains to gender, the social construct, rather than biological sex. It was not our intention to exclude people with more fluid, less easily categorizable, gender identities.

TABLE 2 | Example peace data formats.

Single company, single boundary						
"Person A"	**Group ID**	**Difference Boundary**	**Mediating Tech**	**"Person B"**	**Group ID**	**Economic Value**
Sender	Male	**Gender**	Gmail	Recipient	Female	$ 0.05

This lets you measure the health of your own corporate culture and identify invisible costs and frictions.

Many companies, single boundary						
"Person A"	**Group ID**	**Difference Boundary**	**Mediating Tech**	**"Person B"**	**Group ID**	**Economic Value**
Sender	Female	**Gender**	Gmail	Recipient	Male	$ 0.07
Sender	Male	**Gender**	Facebook @Work	Recipient	Female	$ 0.19
Sender	Female	**Gender**	Salesforce Chatter	Recipient	Male	unknown

This illustrates how data on the same difference boundary can be aggregated across mediating technologies. This lets an industry or geographic region (e.g., the software industry, Silicon Valley, a nation, or the world) see if they are really making progress on a particular issue, if their "solutions" are really working, and quickly identify best practices. Enabling this crucial kind of Peace Data aggregation is one of the key reasons we need the Peace Data Standard proposed in this paper.

Many companies, many boundaries						
"Person A"	**Group ID**	**Difference Boundary**	**Mediating Tech**	**"Person B"**	**Group ID**	**Economic Value**
Sender	Male	**Gender**	Gmail	Recipient	Female	$ 0.08
Host	Gen X	**Age**	Airbnb	Guest	Millennial	$ 0.02
Sender	Pakistani	**Nationality**	Western Union	Recipient	Male	unknown
Driver	West Palo Alto	**Neighborhood**	Uber app	Rider	East Palo Alto	$ 0.13

This could be the first step towards building a better world across every difference boundary we can detect.

Challenge: increase positive engagement between:	**"Person A"**	**Group ID**	**Difference Boundary**	**Mediating Tech**	**"Person B"**	**Group ID**	**Economic Value**
Other Peace Data dimensions							
Women and men in the workplace	Sender	Male	*Gender*	Gmail	Recipient	Female	$0.06
	Sender	Male	**Gender**	Facebook @Work	Recipient	Female	unknown
Countries	Driver	Pakistani	**Nationality**	Uber app	Rider	Indian	$ 0.04
	Host	Indian	**Nationality**	AIrBnB app	Guest	Pakistani	$ 0.09
	CMO	Pakistani, Female	**Nationality and Gender**	Asana project management app	CTO	Indian, Male	unknown
LGBT Community	Seller	Gay	**Sexual Orientation**	eBay	Buyer	Straight	$ 0.17
	Sender	Straight	**Sexual Orientation**	Paypal	Recipient	Gay	$ 0.06

Once core Peace Data is identified, many other dimensions of interest can either be correlated, or inferred, creating new metadata, often of increasing utility and value.

recommend the following process to determine the level of analysis. First, the point at which an *engagement episode* begins must be established (e.g., the first data point connecting A and B: when a male employee texts a female co-worker). Second, once an episode is reciprocated, it becomes an *interaction* (e.g., the female co-worker replies). Third, over time, a series of interactions accrue and can reveal novel, quantitative aspects of the *relationship* between A and B (e.g., A averages 39.5 min latency in response to messages from B). Fourth, an aggregation of relationships between others who share a group identity of interest with A and B (in this case other men and women in this or other workplaces), reveals previously invisible *group dynamics* about those groups of interest, and about the connections between them (e.g., finding that in company X, men typically take 45.75 min to respond to messages from female co-workers, whereas in company Y they take 32.5 min (resulting in an annual

difference of x in productivity, in which x can now potentially be quantified using internal corporate data[7]). These are the kinds of insights that can then inform effective interventions, research, and software design.

A Social Network Analysis Example

We recently collected real world data consistent with our proposed Data Standard (Guadagno et al., 2018). Thus, while the two previous examples presented in **Table 1** were hypothetical in nature, this example comes from a social network analysis of data on employee's social media use an Australian bank. In

---

[7]None of the authors of the present paper have the appropriate academic credentials (e.g., none of us are, for instance, a behavioral economist) so linking economic value to Peace Data is beyond the scope of the present paper. Nonetheless, we think the potential to link financial outcomes is there and are on the lookout for collaborators to help further explore this research direction.

Western society, there is a commonly held stereotype that women talk more than do men. However, empirical investigations of this question have resulted in a different conclusion, one that suggests that the gender difference is in the opposite direction. For instance, the results of two meta-analyses that examined the question found that men generally talk more than do women (James and Drakich, 1993; Leaper and Ayres, 2007). Similarly, a study in which men and women's daily conversations were recorded also concluded that the notion that women talk more than men is a negative stereotype with no real-world basis (Mehl et al., 2007). It is notable that these three research examples pertain to data that were collected long before the widespread adoption of Web 2.0 technologies such as social media.

With respect to social media use, myriad studies have demonstrated that men and women differ in the reasons they use the platform. For instance, research has shown that women generally use social media for relationship maintenance and social comparison purposes, while men use social media to make new connections (Haferkamp et al., 2012; Muscanell and Guadagno, 2012). What we currently do not know is whether men and women's networking patterns change when the context changes to the workplace; and especially the larger corporate environments, that have been traditionally dominated by male executives, in command and control work structures.

To test the question of gender differences in the use of social media, the researchers measured the online social networking patterns of a large financial institution over a period of 6 months. Anonymized data from the online collaboration platform (Yammer – a Microsoft-owned social media platform for the workplace[8]) tracked participants contributions overtime. The only group identification collected was participant gender. Data were collected from over 7,500 employees with an approximately 50/50 gender split. The results of this research were stark: Women were more collaborative and communicative networkers than men, on most dimensions. For instance, across 23 collaborative measures (see **Table 3**), 12 showed a statistically significant difference between men and women in their technology-mediated social interactions at work (see **Figure 3**). As the results indicate, the women lead on all dimensions with a significant gender difference. However, it should be noted that one of the dimensions – %Broadcaster – a higher score is interpreted as having a negative connotation owing to the lack of reciprocal interaction reflected in a high score, and therefore provides the only dimension that men outperform women (Guadagno et al., 2018). **Table 3** provides a brief description of the dimensions used and **Figure 3** displays the standardized gender difference for each dimension on which men and women differed. Our analyses also revealed that women reach out to male colleagues far more often (58%) than men reach out to women (33%) co-workers (see **Figure 4**).

Next, we examined the communication that occurs within and across difference boundaries. These results also revealed that women had denser and more reciprocal communication networks relative to their male colleagues (see **Figure 5**). Thus, while the data reported by Guadagno et al. (2018) clearly

---

[8]https://products.office.com/en-us/yammer/yammer-overview

TABLE 3 | Yammer communication dimensions.

Dimension	Description
Posts*	Average number of posts made
Replies Made	Average number of replies made
Replies Received	Average number of replies received
Likes Made*	Average number of likes made
Likes Received	Average number of likes received
Mentions Made*	Average number of mentions made
Mentions Received	Average number of mentions received
Notifies Made*	Average number of notifications made
Notifies Received	Average number of notifications received
Give Receive*	Balance of Giving (outward) minus Receiving (inward)
Interactions*	Average number of total interactions
Connections*	Average number of unique connections
2-Way Connections*	Average number of two-way connections
Replies/Post	Average number of replies received for each post
Reciprocity	Proportion of connections that are two-way (reciprocated)
In-Connections	Average number of inward connections (e.g., people who have replied to you)
Out-Connections	Average number of outward connections (e.g., people who you have replied to)
Diversity	Average breadth of Yammer groups actively participated in
%Participation*	Average % of those active more than once every 2 weeks (non-observers)
%Engager	Average % Engagers (have a balance between giving and receiving)
%Catalyst	Average %Catalysts (receive more than they give)
%Responder*	Average %Responder (give more than they receive)
%Broadcaster*	Average %Broadcaster (post more but receive less responses)

The variables with asterisks next to them indicate that women were significantly higher than men on that particular dimension.

demonstrates meaningful differences in behavior across the difference boundary of gender. As discussed below, these results can contribute to our understanding of the role of mediated communication in facilitating prosocial behavior.

The Australian banking data uses gender as the relevant group identity (Guadagno et al., 2018) and we suggest that this aspect of our data illustrates the first dimension of Peace Data. One potential extension of this research would be to look for subgroups within gender. For instance, we could compare participants from the Australian banking study that did not differentiate between men and women in their messaging behavior from those that did as a first step in understanding what aspects of group identity (in addition to gender) determine who people are likely to message. Furthermore, illustrative of the second dimension of Peace Data, our Australian banking data example uses the actual behavior of men and women in

FIGURE 3 | This graph displays standardized gender differences for the 12 communication dimensions in which women outperformed men. These bars compare the relative, standardized gender difference on the communication dimensions in the Guadagno et al.'s (2018) social network analysis. Maxed out bars indicate the group that significantly outperformed the other group.

FIGURE 4 | Cross-gender collaboration. This image indicates the percentage of participants by gender in Guadagno et al.'s (2018) social network analysis who sent messages to men vs. women.

the organization to assess the extent to which men and women communicate to others in their workplace, with women engaging in far more communication relative to men. Similarly, given the

longitudinal nature of this data, it is also illustrative of the third dimension of Peace Data. Finally, a great deal of metadata – our fourth dimension of Peace Data – can be extrapolated and analyzed from underlying data; for example, message timestamp *data* in our Australian bank example above allows the calculation of response latency *metadata*. This means much progress can be made by discovering novel metadata calculation formulas (Guadagno et al., 2018). The reciprocal aspect of the data was determined from metadata recorded along with the content of the messages. Finally, what is unknown with this data is the emotional valence of the message content and how this plus the greater communicativeness of women over men relate to economic outcomes such as salary and promotions.

GENERAL DISCUSSION

Mediating technologies, like many technologies, are a double-edged sword owing largely to the unanticipated positive and negative consequences of technology. While they can connect geographically distributed people from all facets of people's lives, much emphasis has been placed on the downsides of this connectedness (e.g., negative social comparison, viral spread of disinformation, divisive, uncivil discourse; Wiederhold, 2016). Since many of these negative effects are likely to be pursued by bad actors (such as trolls, social engineers, hackers) no matter what, we suggest that these same technologies need to
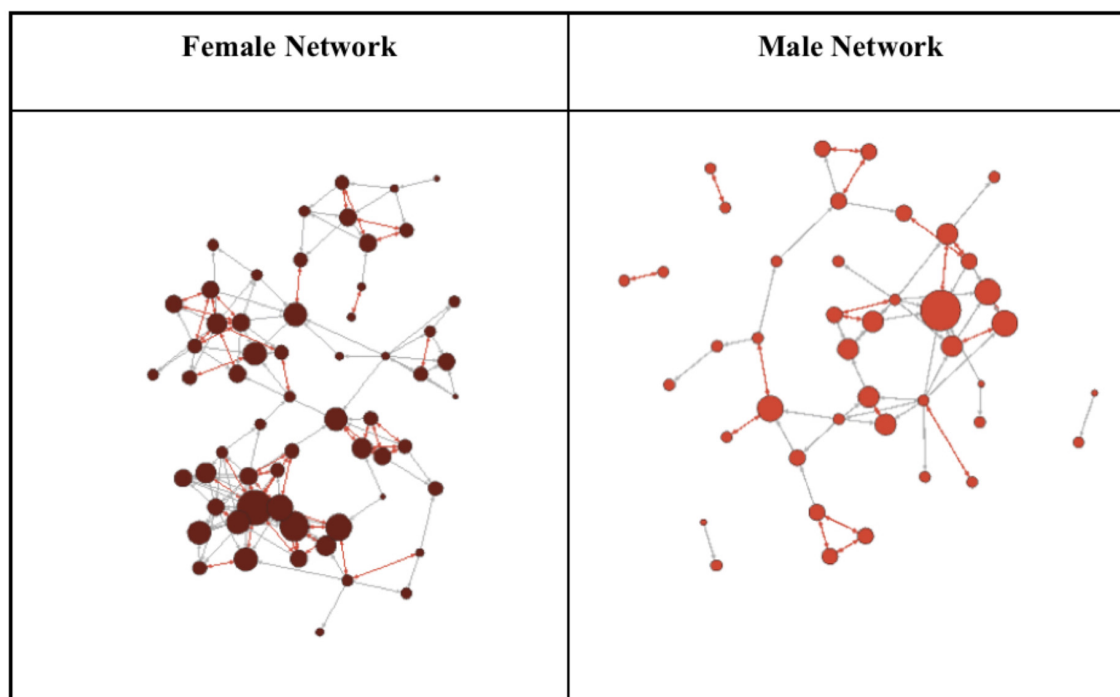
FIGURE 5 | The following diagram compares the female only network with the male only network from Guadagno et al.'s (2018) social network analysis example. The red lines show reciprocated relationships, the gray lines show relationships, and the red dots indicate network nodes with larger nodes indicative of more connections and therefore more influence with the communication network.

also be deliberately designed, used, and cultivated to increase peace, by measurably increasing positive engagement across difference boundaries. Our framework can be applied to a wide variety of issues that organizations and the world at large are currently grappling with, such as increasing inclusivity for women and other members of underrepresented groups in the software industry and the military. As much research demonstrates, diversity of thought and perspective enhances innovation (Nielsen et al., 2017), therefore facilitating positive engagement between people from traditionally underrepresented groups and people in the traditionally dominant group for an occupational category (e.g., software engineer, air force officer) can enhance innovation and success in organizations.

The research and hypothetical examples presented in our paper illustrate how we can now start to identify a broader variety of group identities, some of which may have been unknown before the era of big data allowed using people's digital footprints (e.g., Kosinski et al., 2016). Group identity data may also vary. As previous research has demonstrated, people have many different social or group identities, and the importance of each on people's behavior is context dependent (e.g., Turner et al., 1994). Thus, the nature of the interaction, the social identities of the others present during the interaction, the setting in which the interaction takes place, all affect the outcomes, and, in the age of Big Data, we suggest that all of these features should be recorded and considered with Peace Data. Finally, in addition to the new, emerging group identities, there is also the question of the group identities created (often though not always unintentionally) by

the people/cases left behind as these new kinds of groups are identified and categorized (Leigh Star, 2010). These *residual categories* are the ones that are discarded as unimportant by the data analyst. Leigh Star (2010) argues that determining who the residual categories are and how they are identified in datasets has important ethical implications for understanding certain instances of people's actions and behaviors.

Peace Data Uses

Why collect Peace Data? As illustrated throughout this paper, we argue that Peace Data is useful for many purposes. For instance, Peace Data can reveal largely unseen, invisible relationships and dynamics within those relationships. This can allow use to precisely measure and predict peace akin to Google Earth for social interaction. Furthermore, a fruitful future direction involves the use of Peace Data to determine the economic value of peace (i.e., what is the economic value of a certain action? For instance, it may be that a bank teller warmly greeting a customer is worth 3.17 cents to banks by maintaining a warm friendly atmosphere in a neighborhood where people know each other by name. While the economic value would vary by context (e.g., differ by community and neighborhood, by time of day), and across the additional difference boundaries such as race, gender, and profession. Thus, it may be possible to determine whether the data itself has economic value in addition to the positive prosocial behavior it signifies; for instance, hypothetically, we could use Peace Data to determine that a warm greeting in a high-crime neighborhood at 2:15 a.m. may be worth $4.33,

while a similar warm greeting in a lower-crime neighborhood at 1 p.m. may only be worth $0.79. The value of having this Peace Data becomes clearer when it is applied to inform designers and engineers of peace technology interventions. Furthermore, we assert that insights gained from Peace Data can also be useful to communities seeking to improve the quality of life, health, and social capital. Finally, we suggest that this knowledge may have unintended negative consequences such as driving away consumers, businesses, and residents in geographic locations which generate less peace-related outcomes. Thus, the ethics regarding both modeling and reporting about the economic Peace Data should be carefully considered to avoid such unanticipated negative outcomes.

Ethical Considerations

The proposed research agenda presented in this manuscript is not without serious ethical concerns pertaining to participant consent, security of participant data, and participant anonymity. Consistent with this, Gosling and Mason (2015) noted that one of the ethical issues associated to the study of mediated communication is that: "researchers have less control over and knowledge of the research environment and cannot monitor the experience of participants, or indeed their true identities, raise a number of ethical issues (Buchanan and Williams, 2010)" (p. 894). Thus, we recommend that companies and scholars interested in adopting our proposed Peace Data Standard adopt practices the protect their participants, particularly when they are customers as well.

With respect to participant anonymity, Chandler and Shapiro (2016) point to issues concerning participant anonymity as one of the main challenges of this area of research. For instance, Dawson (2014) provided evidence outside parties can sometimes identify participants whose text-responses were collected via networked applications. If an interested party has the right expertise and/or uses data triangulation methods such as a Google search, the author provided evidence that it is possible to identify participants despite the anonymization of the data. Specifically, Dawson (2014) was able to identify the source of a text passage directly quoted in 10 of 112 articles identified as relevant to this study. Furthermore, of these 10 articles containing identifiable data, the authors of five articles neither anonymized the text nor discussed ethical considerations, and the authors of one article tried unsuccessfully to anonymize the data. Thus, while it is important to safeguard participants' anonymity and confidentiality, more may be required in terms of data obfuscation to protect people's data.

Gosling and Mason (2015) further note that many of the ethical challenges of collecting data from people's networked application use has arisen from a lack of guidance from ethics organizations because the rules developed for the ethical use of human research participants were developed before research on the Internet became ubiquitous. With respect to participant privacy, these authors suggested the criteria for defining behavior as "public" and therefore open for use in research without participant consent, should be carefully considered. They assert that this of particular concern when researchers, both in industry and academia, use web scraping techniques to collect data from Internet-comment and discussion forums such as social media, blogs, and other online discussions.

Similarly, as people's online behavior grows and changes over time, people unintentionally build digital dossiers – a file or set of files containing detailed records about a person's activities on the Internet – on themselves based their Internet use (Guadagno, in press). This accumulation of data has been further enhanced by services provided by social media companies (e.g., Facebook and LinkedIn) that provide the option for people to use their social media user ID and password in to access third party websites. This allows people's online activities to be tracked across different Internet venues and provide useful data on people's technology use, but it comes at a cost to people's online privacy. While we suggest above that rich and detailed Peace Data could be collected by tracking people's application use across websites and services, this should not occur at the cost of participants' privacy and anonymity. As far as we can ascertain, there are no current widely adopted guidelines among researchers on this issue. However, we recommend the insightful and detailed questions to consider proposed by Buchanan and Williams (2010) pertaining to ethical considerations such as how to determine whether a specific Internet behavior is "public" and can be recorded and/or observed without consent. Finally, while many corporate entities fold blanket consent to participate in research as part of end user licensing agreements, we further suggest that industry researchers consider crafting more direct and more educational consent procedures to both inform and protect the people whose data they collect.

Limitations and Future Directions for Research

We are in the nascence of technological growth especially with applications that leverage large scale group dynamics inherent online (e.g., crowdsourcing applications, review aggregators, social media). While there are obvious unintended negative consequences (i.e., fake news spreads rapidly unchecked, the divisive direction of online discussions often take, especially if the topic is controversial) of this rapid technology growth, in this paper, we questioned if it is possible to turn this rapid technological growth into something positive? The Peace Data Standard presented in this paper is a first step in leveraging the power of Big Data and machine learning to start to understand when and how technology facilitates peace and positivity – people being good to each other. This paper is written as a conceptual paper and, as such, we did not include a discussion of the different database, programming, and data analysis tools that are currently popular with people who work with big data (e.g., R, Python, JavaScript, and SQL) as these tools may change over time.

The Peace Data Standard presented in this manuscript suggests several clear future directions for research. First, we can begin to identify different subgroups with increasing resolution and precision, particularly based on behavior sequences (which we refer to as engagement episodes); categorizing and correlating prosocial behaviors with increased precision. This correlation between behaviors and/or sequences of behaviors to outcomes

of interest will generally occur longitudinally. Furthermore, we suggest that scholars in industry and academia with backgrounds in fields such as behavioral economics take up our call to build statistical models to understand the economic benefits of Peace Data. Similarly identifying the kinds of additional metadata useful to predict outcomes related to understanding the role of technology in facilitating peace is also an important for understanding the underlying mechanisms that promote positive engagement episodes resulting in peaceful outcomes.

Additionally, while many corporate entities likely already examine this internally, we would like to call for future research to examine the economic aspects of the different behaviors peace-relevant we can track using Big Data. Finally, given the paucity of research on the interplay between relationships that form online then transition to offline contexts, we would like to make a call for more research on the ways in which contact initiated online and/or used in context with offline interactions affects interpersonal processes.

An Invitation

In conclusion, we see many opportunities for contribution from experts across many disciplines:

1. *Help improve the standard.* We very intentionally propose this as the version 0.1 beta of the Peace Data Standard. It should be dynamic, and there is much room for improvement from many intersecting domains. We therefore invite colleagues from academia, industry, and government around the world to respond to this paper with recommendations for further refinements for future release versions of the Peace Data Standard. These refinements should reflect, but not be limited to, psychological, sociological, technological, ethical, and methodological advancements in data collection and analysis from large-scale, real-time mediated technology. This includes taking advantage of increasing capability in the underlying sensor and computation technologies that make Peace Data collection and analysis possible. We propose working together toward establishing version 1.0 as a formal ISO standard for Peace Data. This is vital to enabling the kind of comparative analysis across platforms, organizations, communities, and locations, that will yield the fastest and cheapest discovery of what works best, where, when, and under what conditions, thus providing the greatest benefit for all.

2. *Help source Peace Data.* Colleagues in academia can teach students to help local businesses identify and submit their Peace Data. Colleagues in industry can spread the word internally and in their sector about the value of a company's Peace Data, for things like regulator relations, customer engagement and loyalty, talent recruitment and retention, and more.

3. *Help make Peace Data auditable, so it becomes an effective market signal for (at least some of) the true value peace creates.* Organizations with Peace Data must be able to guarantee the validity of their data to preserve its value, and organizations and individuals relying on that data

must be able to ensure it has not been tampered with or inflated. Therefore, third party audit standards and processes need to be developed jointly by all stakeholders, to ensure reliability and preserve meaning. Given the huge scale, granularity, and real-time nature of Peace Data, there is also a need for these audit processes to be automated as much as possible.

We further invite readers to consider whether this framework for Peace Data is applicable to the type of data collected in their research and by their organizations. While it is becoming more common in industry to hire data scientists (De Mauro et al., 2017), we suggest that data scientists can be particularly helpful in employing our proposed Peace Data Standard. Specifically, data scientists can be enlisted to help identify data that could be used to promote peaceful interactions through the use of mediated technology, sitting in the epicenter of Silicon Valley we understand firsthand the scarcity of good data scientists.

4. *Help establish a Peace Data Prize.* Having an established Peace Data Standard will enable companies who generate Peace Data to show their audited peace impact to regulators, employees, and customers. The public relations benefit alone, not to mention the improved customer loyalty and engagement, or the talent recruitment and retention benefits, may enable member companies who submit their Peace Data for audit to pool a small percentage of that aggregate new value, and fund a series of Peace Data Prizes, which the Peace Innovation Lab (PIL) at Stanford[9] will award annually to customers, employees, and companies who have had the greatest per capita peace impact, who have made the greatest peace improvements, and so forth. This same arrangement can be repeated for municipalities, national governments, civil and religious organizations, and so forth. We propose to also set aside a portion of these funds for research, and for prizes for the best research using Peace Data each year. Interested organizations, corporate or government, should contact the PIL to register their participation.

Implications

For industry, this Peace Data Standard facilitates a results-based economic mechanism for stakeholders to invest directly in peace. We argue that this enables a new kind of precision peace, in scenarios such as follows: first, from the data we presented above, any company wishing to understand gender differences in workplace behavior, can now measure within- and cross-gender engagement in the workplace, to identify whether there are any problems with gender discrimination, and, if so, what interventions actually work in their context. Second, for a US bank that needs to meet their Community Reinvestment Act (CRA) compliance requirements, being able to post a results-based contract for any entrepreneur who can design technology interventions that can elicit positive prosocial economic behaviors in their underprivileged CRA district. Third,

---

[9]The academic home of the first and second authors of this paper.

for a city government whose tax base depends largely on property taxes, being able to pay for precision targeted positive engagement that increases quality of life – and thus property values and property taxes – in any neighborhood in their city. Fourth, for a municipal-bond underwriter being able to insure repayment of their bond by investing in exactly the same scenario as the city government, in the example above.

Implications of the Peace Data Standard for academia suggest its adoption can facilitate the following research avenues: rigorous empirical examination of first and second order economic impacts with types and qualities of engagement episodes mediated by different technologies; rapid, large N hypothesis testing for psychological and sociological impacts of technology platforms; deployment of large scale randomized controlled field trials to validate promising hypotheses in many different environments and under many varying conditions; testing and deployment of machine learning to generate automated rapid responses to changing conditions anywhere in the world. As stated earlier in the manuscript, we believe that this data standard will also allow scholars and entrepreneurs to rapidly design, test, and validate Peace Tech interventions that effectively transform these group differences into raw material for sustainable peace (in which mutual benefit is equal to the cost of engagement) and eventually scalable positive peace (in which mutual benefit exceeds the cost of engagement).

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Allport, G. W. (1954/1979). *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Amichai-Hamburger, Y. (2005). Internet minimal group paradigm. *Cyberpsychol. Behav.* 8, 140–142. doi: 10.1089/cpb.2005.8.140

Amichai-Hamburger, Y. (2015). Online conflict reduction. *Comput. Hum. Behav.* 52:507. doi: 10.1016/j.chb.2015.03.002

Amichai-Hamburger, Y., Hasler, B. S., and Shani-Sherman, T. (2015). Structured and unstructured intergroup contact in the digital age. *Comput. Hum. Behav.* 52, 515–522. doi: 10.1016/j.chb.2015.02.022

Amichai-Hamburger, Y., and McKenna, K. Y. A. (2006). The contact hypothesis reconsidered: interacting via the internet. *J. Comput. Mediat. Commun.* 11, 825–843. doi: 10.1111/j.1083-6101.2006.00037.x

Anderson, R. (2004). A definition of peace. *Peace Conf.* 10, 101–116. doi: 10.1207/s15327949pac1002_2

Azucar, D., Marengo, D., and Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: a meta-analysis. *Pers. Individ. Dif.* 124, 150–159. doi: 10.1016/j.paid.2017.12.018

Ball, A. (2017). *Metadata: The Spirit of Research Data Management? In: Research Infrastructures in Social Sciences and Humanities, 2017-06-29.* Available at: http://opus.bath.ac.uk/56428/

Balliet, D., Wu, J., and De Dreu, C. K. (2014). Ingroup favoritism in cooperation: a meta-analysis. *Psychol. Bull.* 140, 1556–1581. doi: 10.1037/a0037737

Baumeister, R. F., Vohs, K. D., and Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: whatever happened to actual behavior? *Perspect. Psychol. Sci.* 2, 396–403. doi: 10.1111/j.1745-6916.2007.00051.x

Brewer, M. B. (1988). "A dual process model of impression formation. advances in social cognition," in *A Dual Process Model of Impression Formation*, Vol. 1, eds T. K. Srull and R. S. Wyer (Hillsdale, NJ: Lawrence Erlbaum Associates), 1–36.

Brewer, M. B. (1991). The social self: on being the same and different at the same time. *Pers. Soc. Psychol. Bull.* 17, 475–482. doi: 10.1177/0146167291175001

Buchanan, T., and Williams, J. E. (2010). "Ethical issues in psychological research on the Internet," in *Advanced Methods for Conducting Online Behavioral Research*, eds S. D. Gosling and J. A. Johnson (Washington, DC: American Psychological Association), 255–271.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

burrough, x. (2016). Meditations on a digital workforce. *Leonardo* 49, 436–437. doi: 10.1162/LEON_a_01292

Cao, B., and Lin, W. Y. (2017). Revisiting the contact hypothesis: effects of different modes of computer-mediated communication on intergroup relationships. *Int. J. Intercult. Relat.* 58, 23–30. doi: 10.1016/j.ijintrel.2017.03.003

Chandler, J., and Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psychol.* 12, 53–81. doi: 10.1146/annurev-clinpsy-021815-093623

Choi, H., and Varian, H. (2012). Predicting the present with Google Trends. *Econ. Rec.* 88, 2–9. doi: 10.1111/j.1475-4932.2012.00809.x

Christie, D. J., and Montiel, C. J. (2013). Contributions of psychology to war and peace. *Am. Psychol.* 68, 502–513. doi: 10.1037/a0032875

Cohrs, J. C., Christie, D. J., White, M. P., and Das, C. (2013). Contributions of positive psychology to peace: toward global well-being and resilience. *Am. Psychol.* 68, 590–600. doi: 10.1037/a0032089

Crocker, J., Major, B., and Steele, C. (1998). "Social stigma," in *The Handbook of Social Psychology*, 4th Edn, Vol. 2, eds D. Gilbert, S. T. Fiske, and G. Lindzey (Boston, MA: McGraw-Hill), 504–553.

Dawson, P. (2014). Our anonymous online research participants are not always anonymous: is this a problem? *Br. J. Educ. Technol.* 45, 428–437. doi: 10.1111/bjet.12144

De Mauro, A., Greco, M., Grimaldi, M., and Ritala, P. (2017). Human resources for Big Data professions: a systematic classification of job roles and required skill sets. *Inf. Proc. Manag.* (in press). doi: 10.1016/j.ipm.2017.05.004

Deloitte Access Economics (2016). *Economic Effects of Ridesharing in Australia.* Available at: https://www2.deloitte.com/content/dam/Deloitte/au/Documents/Economics/deloitte-au-economics-economic-effects-of-ridesharing-australia-150216.pdf

Galtung, J. (1969). Violence, peace, and peace research. *J. Peace Res.* 6, 167–191. doi: 10.1177/002234336900600301

Goffman, E. (2009). *Stigma: Notes on the Management of Spoiled Identity*. New York, NY: Simon and Schuster.

Gore, M. (1984). *People are People. On Some Great Reward*. Berlin: Mute Records.

Gosling, S. D., and Mason, W. (2015). Internet research in psychology. *Annu. Rev. Psychol.* 66, 877–902. doi: 10.1146/annurev-psych-010814-015321

Guadagno, R. E. (in press). "Using the internet for research," in *Advanced Research Methods for the Social Sciences*, eds J. Edlund and A. L. Nichols (Cambridge: Cambridge University Press).

Guadagno, R. E., Lock Lee, L., Kjaer, C., and Nelson, M. (2018). Understanding gender bias in the workplace: a 6-month longitudinal study of people's messaging practices. Manuscript in preparation.

Haferkamp, N., Eimler, S. C., Papadakis, A. M., and Kruck, J. V. (2012). Men are from Mars, women are from Venus? Examining gender differences in self-presentation on social networking sites. *Cyberpsychol. Behav. Soc. Netw.* 15, 91–98. doi: 10.1089/cyber.2011.0151

Hogg, M. A. (2016). "Social identity theory," in *Understanding Peace and Conflict Through Social Identity Theory*, eds S. McKeown, R. Haji, and N. Ferguson (Cham: Springer International Publishing), 3–17.

James, D., and Drakich, J. (1993). "Understanding gender differences in amount of talk: a critical review of research," in *Oxford Studies in Sociolinguistics. Gender and Conversational Interaction*, ed. D. Tannen (New York, NY: Oxford University Press), 281–312.

Kooti, F., Lerman, K., Aiello, L. M., Grbovic, M., Djuric, N., and Radosavljevic, V. (2016). "Portrait of an online shopper: understanding and predicting consumer behavior," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (New York, NY: ACM), 205–214. doi: 10.1145/2835776.2835831

Kosinski, M., Wang, Y., Lakkaraju, H., and Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychol. Methods* 21, 493–506. doi: 10.1037/met0000105

Leaper, C., and Ayres, M. M. (2007). A meta-analytic review of gender variations in adults' language use: talkativeness, affiliative speech, and assertive speech. *Pers. Soc. Psychol. Rev.* 11, 328–363. doi: 10.1177/1088868307302221

Leigh Star, S. (2010). Residual categories: silence, absence and being an other. *Z. Medien Kulturforschung* 2010, 201–219.

Lewandowski, G. W. Jr., and Strohmetz, D. (2009). Actions can speak as loud as words: measuring behavior in psychological science. *Soc. Personal. Psychol. Compass* 3, 992–1002. doi: 10.1111/j.1751-9004.2009.00229.x

MacInnis, C. C., and Hodson, G. (2015). The development of online cross-group relationships among university students: benefits of earlier (vs. later) disclosure of stigmatized group membership. *J. Soc. Pers. Relat.* 32, 788–809. doi: 10.1177/0265407514548394

Major, B., and O'Brien, L. T. (2005). The social psychology of stigma. *Annu. Rev. Psychol.* 56, 393–421. doi: 10.1146/annurev.psych.56.091103.070137

McKeown, S., and Dixon, J. (2017). The 'contact hypothesis': critical reflections and future directions. *Soc. Personal. Psychol. Compass* 11:e12295. doi: 10.1111/spc3.12295

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., and Pennebaker, J. W. (2007). Are women really more talkative than men? *Science* 317, 82–82.

Muscanell, N. L., and Guadagno, R. E. (2012). Make new friends or keep the old: gender and personality differences in social networking use. *Comput. Hum. Behav.* 28, 107–112. doi: 10.1016/j.chb.2011.08.016

Nielsen, M. W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H. J., Joshi, A., et al. (2017). Opinion: gender diversity leads to better science. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1740–1742. doi: 10.1073/pnas.1700616114

Nowak, A., Vallacher, R., Strawińska, U., and Brée, D. S. (2013). "Dynamical social psychology: an introduction," in *Complex Human Dynamics*, eds A. Nowak, K. Winkowska-Nowak, and D. Brée (Heidelberg: Springer), 1–19. doi: 10.1007/978-3-642-31436-0

Onorato, R. S., and Turner, J. C. (2004). Fluidity in the self-concept: the shift from personal to social identity. *Eur. J. Soc. Psychol.* 34, 257–278. doi: 10.1002/ejsp.195

Peace (n.d.). *In Merriam-Webster*. Available at: https://www.merriam-webster.com/dictionary/peace

Pettigrew, T. F., and Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *J. Pers. Soc. Psychol.* 90, 751–783. doi: 10.1037/0022-3514.90.5.751

Quihuis, M., Nelson, M., and Guttieri, K. (2015). Peace technology: scope, scale, and cautions. *Building Peace* 5, 14–16.

Reeves, B., and Nass, C. (1996). *The Media Equation. Three Paradoxes of the Information Age*. New York, NY: Cambridge University Press.

Sherif, M. (1961). *The Robbers Cave Experiment: Intergroup Conflict and Cooperation*. Middletown, CT: Wesleyan University Press.

Smith, A. (2016). *Shared, Collaborative and On Demand: The New Digital Economy*. Washington, DC: Pew Internet & American Life Project.

Spears, R., Postmes, T., Lea, M., and Wolbert, A. (2002). The power of influence and the influence of power in virtual groups: a SIDE look at CMC and the Internet. *J. Soc. Issues* 58, 91–108. doi: 10.1111/1540-4560.00250

Tajfel, H. (1970). Experiments in intergroup discrimination. *Sci. Am.* 223, 96–102. doi: 10.1038/scientificamerican1170-96

Tajfel, H., and Turner, J. C. (1986). "An integrative theory of intergroup conflict," in *Psychology of Intergroup Relations*, eds S. Worchel and W. Austin (Chicago, IL: Nelson-Hall), 2–24.

Turner, J. C., Oakes, P. J., Haslam, S. A., and McGarty, C. (1994). Self and collective: cognition and social context. *Pers. Soc. Psychol. Bull.* 20, 454–463. doi: 10.1177/0146167294205002

van der Zee, R. (2016). *The 'Airbnb Effect': Is It Real, and What Is It Doing to a City Like Amsterdam? The Guardian*. Available at: https://www.theguardian.com/cities/2016/oct/06/the-airbnb-effect-amsterdam-fairbnb-property-prices-communities

Walther, J. B., Hoter, E., Ganayem, A., and Shonfeld, M. (2015). Computer-mediated communication and the reduction of prejudice: a controlled longitudinal field experiment among Jews and Arabs in Israel. *Comput. Hum. Behav.* 52, 550–558. doi: 10.1016/j.chb.2014.08.004

Wen, Y. T., Yeh, P. W., Tsai, T. H., Peng, W. C., and Shuai, H. H. (2018). "Customer purchase behavior prediction from payment datasets," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, (New York, NY: ACM), 628–636. doi: 10.1145/3159652.3159707

White, F. A., Harvey, L. J., and Abu-Rayya, H. M. (2015). Improving intergroup relations in the Internet age: a critical review. *Rev. Gen. Psychol.* 19, 129–139. doi: 10.1037/gpr0000036

Wiederhold, B. K. (2016). Social media sites, part 1: the dark side. *Cyberpsychol. Behav. Soc. Netw.* 19, 157. doi: 10.1089/cyber.2016.29027.bkw

Conflict of Interest Statement: LLL was employed by company SWOOP Analytics.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Theoretical Modeling of Cognitive Dysfunction in Schizophrenia by Means of Errors and Corresponding Brain Networks

Yuliya Zaytseva [1,2]*, Iveta Fajnerová [1], Boris Dvořáček [1], Eva Bourama [2], Ilektra Stamou [2], Kateřina Šulcová [1,2], Jiří Motýl [1], Jiří Horáček [1,2], Mabel Rodriguez [1] and Filip Španiel [1,2]

[1] National Institute of Mental Health, Klecany, Czechia, [2] 3rd Faculty of Medicine, Charles University in Prague, Prague, Czechia

The current evidence of cognitive disturbances and brain alterations in schizophrenia does not provide the plausible explanation of the underlying mechanisms. Neuropsychological studies outlined the cognitive profile of patients with schizophrenia, that embodied the substantial disturbances in perceptual and motor processes, spatial functions, verbal and non-verbal memory, processing speed and executive functioning. Standardized scoring in the majority of the neurocognitive tests renders the index scores or the achievement indicating the severity of the cognitive impairment rather than the actual performance by means of errors. At the same time, the quantitative evaluation may lead to the situation when two patients with the same index score of the particular cognitive test, demonstrate qualitatively different performances. This may support the view why test paradigms that habitually incorporate different cognitive variables associate weakly, reflecting an ambiguity in the interpretation of noted cognitive constructs. With minor exceptions, cognitive functions are not attributed to the localized activity but eventuate from the coordinated activity in the generally dispersed brain networks. Functional neuroimaging has progressively explored the connectivity in the brain networks in the absence of the specific task and during the task processing. The spatio-temporal fluctuations of the activity of the brain areas detected in the resting state and being highly reproducible in numerous studies, resemble the activation and communication patterns during the task performance. Relatedly, the activation in the specific brain regions oftentimes is attributed to a number of cognitive processes. Given the complex organization of the cognitive functions, it becomes crucial to designate the roles of the brain networks in relation to the specific cognitive functions. One possible approach is to identify the commonalities of the deficits across the number of cognitive tests or, common errors in the various tests and identify their common "denominators" in the brain networks. The qualitative characterization of cognitive performance might be beneficial in addressing diffuse cognitive alterations presumably caused by the dysconnectivity of the distributed brain networks. Therefore, in the review, we use this approach in the description of standardized tests in the scope of potential errors in patients with schizophrenia with a subsequent reference to the brain networks.

Keywords: cognitive deficits, schizophrenia, cognitive tests, errors, brain networks, lesions, fMRI

INTRODUCTION

Cognitive dysfunction in schizophrenia has been propagated as a core component of the illness since Kraepelin (Kraepelin, 1919). Until now, there is no evident comprehension of the mechanisms of cognitive disturbances in schizophrenia. On the one hand, regional brain volume alterations in patients with schizophrenia as compared to healthy individuals, are associated with IQ-dependent cognitive measures, e.i verbal and non-verbal memory, processing speed (for review see Antonova et al., 2005). Though, the patterns of cognitive deficits seem to be more complex that the pattern of structural alterations. On the other hand, the dysconnectivity theory (referred to as "cognitive dysmetria") of schizophrenia suggests that the cognitive deficits might originate from the aberrant functional brain networks activity (Andreasen et al., 1996). This disrupted connectivity results in altered functional integration since it involves either exaggerated connections or weakened pathways (Stephan et al., 2006; Fornito et al., 2011). The deficits in attention and working memory, as it was shown by Whitfield-Gabrieli et al, correlate with the alterations in networks coupling. Specifically, the lack of suppression of the default-mode network (DMN), which intends to suppress during information processing, implies the disbalanced excitatory/inhibitory brain circuits in schizophrenia (Whitfield-Gabrieli et al., 2009).

For the successful information processing, well-coordinated functioning of the distinct brain structures is essential (Pöppel, 1989). Neurological, neuropsychological and neuroimaging studies show that genuinely all cognitive functions rely on the enactments of the dispersed cortical and subcortical brain structures and are not restrained to the specific structures (Singer, 2013; Sporns, 2013). Presumably, cognitive operations arise out of a composed action in the brain networks. Given the multidimensional organization of cognitive functions, the identification of the neural networks responsible for the specific cognitive functions seems to be problematic, especially since the brain structures are oftentimes associated with a number of cognitive operations. The one-to-one relationship of specific brain region and cognitive function does not seem to hold true to the psychiatric diseases in general and to schizophrenia in particular.

Definition of the Qualitative Characteristics of Cognitive Functioning in Patients With Schizophrenia

Neuropsychological studies of schizophrenia have a long history. The functional alterations are pronounced in motor and perceptual processes, spatial functions, verbal and non-verbal memory, attention and executive functioning (Green and Harvey, 2014). Consistent results across the studies demonstrated that if not all, the majority of schizophrenia patients perform more poorly than healthy controls (Mesholam-Gately et al., 2009). In accordance with Harvey (Harvey, 1997), the depth of the deficit can be described relatively to the corresponding reduction of performance in the number of SDs compared to the population norm and the number of affected functions as following: "mild," 0.5–1.0 SD, perceptual

capacity, remote memory; "moderate," 1.0–2.0, remote, short and working memory, attention, and visuomotor functions; and "severe cognitive disability," 2.0–5.0 SD, learning, executive function, memory, vigilance, motor functions, verbal fluency). Subgroups of individuals with SZ may cluster together according to their pattern of cognitive deficits, suggesting the existence of subtypes of dysfunction (Rodriguez et al., 2015). Previous findings agree on two extreme clusters, characterized by near-normal performance on one side (e.g., Goldstein et al., 1998) and profound global dysfunction on the other side (e.g., Goldstein and Shemansky, 1995). One or two remaining subsets are in agreement with the partial deficit (e.g., visual memory and processing speed in Gilbert et al., 2014) with mild cognitive impairment. Standardized scoring in the majority of the neurocognitive tests renders the index scores or the achievement indicating the severity of the cognitive impairment rather than the actual performance. Though such approach facilitates the tracking of the cognitive functioning during the follow-up, at the same time it may lead to the situation when two patients with the same index score of the particular cognitive test demonstrate qualitatively different performances. This may support the view that test paradigms that habitually incorporate different cognitive variables associate weakly, reflecting an ambiguity in the interpretation of noted cognitive constructs (Poldrack, 2011).

Therefore, qualitative approach, being focused on the performances, allows to characterize the types of errors and track of them throughout the testing procedure (Zaytseva et al., 2015). The commonality of the errors across the number of tests may identify the impaired brain structures or networks. Qualitative analysis of tests with the definition of errors (Golden et al., 2000; Strauss et al., 2006) is widely applied in patients with mild cognitive impairment or neurodegenerative diseases (for instance, see Collie and Maruff, 2002; Thompson et al., 2005) that are also known to present with generalized cognitive deficits. The qualitative characterization of cognitive performance might be beneficial in addressing diffuse cognitive alterations presumably caused by the dysconnectivity of the distributed brain networks. The present selective review is focused on the description of the errors in widely used cognitive tests and highlighting the specificities of the performance in schizophrenia patients, hence providing a reference frame for the search of the underlying neural mechanisms.

TRAIL MAKING TEST (TMT)[1]

Behavioral Performance (Errors) and Brain Correlates of TMT in Healthy and Lesion Cohorts

The trail making task includes two variants: TMT-A examines mainly visuoperceptual abilities and processing speed, TMT-B reflects working memory as well as task-switching ability. Two parameters, the time required to complete the test and the number of errors are typically used to measure

---

[1]The Trail Making Test (TMT) is divided into two parts, in part A (TMT-A), typically the subject connects 25 encircled numbers randomly scattered on a page

performance in TMT (Lezak et al., 2012). Subtraction B-A or ratio B/A of the completion time are used to minimize visuoperceptual and working memory demands, thus specifically evaluating mental flexibility and executive control (Sánchez-Cubillo et al., 2009). A study in healthy elderly subjects (Oosterman et al., 2010) supported this functional construction of the TMT scores showing that the predictive value of individual neuropsychological test scores (working memory, executive function, speed and attention, episodic memory) differed among the various TMT-B variables. While the TMT-B total completion time was associated with all neuropsychological scores, only executive function predicted the ratio score (TMT-B/A). In terms of qualitative analysis, the following general categories of observed errors in TMT are omission, perseveration, repetition, sequential and proximity errors (Lezak et al., 2012). **Omission errors** apply mostly to TMT-A and refer to skipping a number in the sequence. **Perseveration errors**, referred to as set-shifting errors, are only seen in TMT-B, when the subject fails to change a set from number to letter and vice versa. **Repetition errors** are made when the same circle is selected more than once. **Sequential errors** can be seen in both TMT variants when the number or letter sequence is incorrect. **Spatial or proximity errors**, also known as capture responses, represent errors in the sequence which occur when the subject circles an incorrect number or letter that is located nearby. Most TMT variants are designed in such a way that pulls for such **proximity errors** (Lezak et al., 2012). Research has shown that normal control subjects can make at least one error on both parts of the TMT (Ruffolo et al., 2000; Lezak et al., 2012) In fact, several factors such as age (Płotek et al., 2014) or educational level (Płotek et al., 2014) and even shift work history (Titova et al., 2016) can affect performance. However, an increased number of errors, especially on TMT-B, has been associated with dorsolateral frontal lobe lesions (Kopp et al., 2015) and this finding has been consistent even when compared to subjects with inferior medial frontal lobe and posterior lobe lesions, who made fewer errors or were completely unaffected (Stuss et al., 2001; Lezak et al., 2012). Chan et al. (2015) challenged the notion that increased number of errors on TMT-B was specific to frontal lesions but did confirm that there was a significant difference in a number of errors when comparing groups with either frontal or non-frontal lesions to healthy controls. Moreover, medial temporal lobe atrophy has been shown to be the strongest neuroanatomical predictor of TMT-B performance in elderly subjects (Oosterman et al., 2010), when analyzed together with periventricular and white matter MRI hyperintensities.

Several variants of the TMT task have been applied as fMRI paradigms. Verbal variant (counting from 1 to 24 or alternate numbers and letters), showed brain activations predominantly in the left frontal lobe hemisphere structures including dorsolateral prefrontal cortex (DLPFC), ventralateral prefrontal cortex (VLPFC) and premotor and motor areas, and

in the right hemisphere the cingulate and intraparietal sulci (Moll et al., 2002).

Zakzanis et al. (2005) used MRI compatible writing device called virtual stylus to perform the original visuospatial TMT task. The study elicited the brain activity in three clusters. The first cluster involved activity in frontal lobe areas of the left hemisphere as reported previously by Moll et al. (Moll et al., 2002), including the medial and dorsolateral prefrontal cortex (PFC), precentral gyrus, cingulate gyrus, and insula. The second cluster in the right hemisphere included cingulate cortex and insula. The smallest third cluster included the left middle and superior temporal gyrus suggesting the utilization of internal speech processing. Similar findings suggesting the involvement of PFC, visuomotor and speech processing areas, including Broca's area, have been reported also in functional near-infrared spectroscopy (fNIS) study by Hagen et al. (Hagen et al., 2014). Another TMT adaptation with fMRI event-related design (Allen et al., 2011) was used in healthy participants, who had to visually scan image with a pseudo-randomly distributed array of 22 items and press the button when correct letter/number was localized. Interestingly, this study failed to find PFC activation previously reported in mental set switching. The results highlighted bilateral activations in ventral and dorsal visual streaming and motor response related brain areas.

Few studies have demonstrated the difference between TMT B and A variants (Moll et al., 2002; Jacobson et al., 2011), showing that TMT-B task elicited stronger brain activity in the bilateral DLPFC, right VLPFC and precentral gyrus and left temporoparietal area. Similarly, combined usage of computer version of TMT and fNIS reported blood flow increases in the bilateral PFC when contrasting TMT B vs. TMT A (Kubo et al., 2008).

The number of studies comparing resting state fMRI (rsfMRI) activity and neuropsychological tests performance is sparse. A recent study in healthy volunteers (James et al., 2016) showed a significant effect of the TMT B performance (time to complete) to resting-state connectivity of two regions of interest (ROIs), the right VLPFC and left superior parietal lobule.

The summary of the brain activations is schematically depicted in **Figure 1**.

Behavioral TMT Performance and Brain Correlates in Schizophrenia Patients

Although errors on the TMT are not specific to one psychiatric condition (Moritz et al., 2001), there is evidence that schizophrenia patients show an overall slower processing speed, impaired visuomotor tracking and switching ability (Rodriguez et al., 2015). The studies focusing on the specific TMT errors in schizophrenia are limited, though they have demonstrated that patients with schizophrenia tend to make some errors more consistently. Notably, Mahurin et al. (2006) showed that schizophrenia patients make more sequencing, or as operationally defined in this research "tracking," errors compared to

---

in ascending order by drawing a pencil line. In part B (TMT-B) there are 25 encircled numbers and letters that should be connected in alternating order.
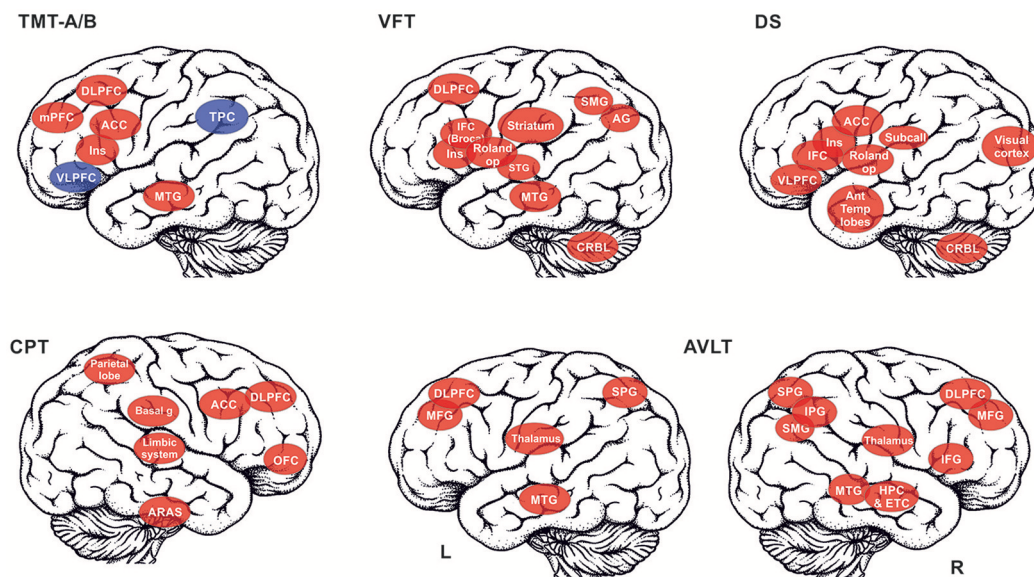
FIGURE 1 | The summary of the brain structures that are recruited in the selected cognitive tests is depicted: TMT A/B, Trail making test, A and B versions (red color for TMT A, red+blue for TMT B); VFT, Verbal fluency test; DS, Digit span; CPT, Continuous performance task; AVLT, Auditory verbal learning test (reported separately for left (L) and right (R) hemisphere). List of abbreviations of brain areas in alphabetical order: ACC, anterior cingulate cortex; AG, angular gyrus, Ant Temp lobes – anterior temporal lobes; ARAS, ascending reticular activation system; Basal g, basal ganglia; DLPFC, dorsolateral prefrontal cortex; CRBL, cerebellum; ETC, entorhinal cortex; HPC, Hippocampus; IFC, interior frontal cortex (e.g., Broca's area); IPG, inferior parietal gyrus; MFG, middle frontal gyrus; mPFC, medial prefrontal cortex; MTG, middle temporal gyrus; OFC, orbitofrontal cortex; Roland op, Rolandic operculum; SMG, supramarginal gyrus; SPG, superior parietal gyrus; STG, superior temporal gyrus; Striatum,– incl. putamen, nucleus caudatus, globus pallidum; Subcall, subcallosum; TPC, temporo-parietal cortex; VLPFC, ventrolateral prefrontal cortex.

both depressed patients and healthy controls, which they attributed to a greater degree of cognitive disorganization in schizophrenia.

Imaging studies using TMT task in schizophrenia are sparse. Some studies reported that TMT performance in schizophrenia patients could be predicted by resting state metabolism measured using positron emission tomography (PET) (Horacek et al., 2006) or by resting state connectivity (Argyelan et al., 2013). A recent study using transcranial direct-current stimulation (tDCS) reported altered hemodynamic pattern during the TMT task performance in middle cerebral arteries (Schuepbach et al., 2016).

CONTINUOUS PERFORMANCE TEST (CPT)[2]

Behavioral Performance (Errors) and Brain Correlates of CPT in Healthy and Lesion Cohorts

In order to test sustained attention and vigilance, many tasks using similar paradigm have been developed. One of the most applied measures is the Continuous performance test (CPT), which assesses four aspects of attention: inattentiveness, impulsivity, sustained attention, and vigilance. The main

[2]The test is administered by PC software and takes approximately 15 min. In standard CPT-X version respondent is required to respond when any letter, except the letter "X," appears on the screen (Conners, 2014).

categories of errors performed in the CPT are omissions and commissions. **Omissions** are made when the respondent does not react to target letters ("non-X"). Results from studies with patients who had damage to the basal ganglia showed more omissions errors (Levin et al., 1986; Wolfe et al., 1990). More omissions errors, longer reaction time and the greatest vigilance decrement are associated with right frontal damage (Rueckert and Grafman, 1996). **Commissions** result from the response to non-target letters ("X"). There are several subtypes of commissions errors: "fast reaction-time response" that is associated with impulsivity and a "slow reaction time response" or "delayed response" as a result of inattention; and a "random" type that is associated with a lack of control (Halperin et al., 1991). **Perseverations** can be also considered as a type of error. The perseverative response in CPT is any reaction time that is less than 100 ms. Such responses are: a slow response to the preceding stimuli, a random response, an anticipatory response, or a response repeated without consideration of the stimuli or task requirements (Conners, 2014).

CPT test is known as a reliable measure of attention (Rosvold et al., 1956; Riccio et al., 2002). Lesion studies show that especially lesions and damages in the right frontal area effect CPT performance. The more severe lesions, the bigger attentional problems, and worse CPT performance has been reported (Katz et al., 1996; Riccio et al., 2002). Riccio et al. (2002) see CPT test as a good symptom-specific measurement but a poor disorder-specific test. They suggested the term

"asymmetry" attention and show in their review that the right hemisphere is more activated during the CPT test and that the test is connected with models of attention including cortical (frontal, temporal, parietal), subcortical [limbic, basal ganglia and ascending reticular activating system (ARAS)] and functional systems (pathways between the basal ganglia, thalamus and frontal lobes) (Riccio et al., 2002). Ogg et al. (2008) show correlation between healthy adults' reaction time in Conners' CPT test and anterior cingulate cortex activation. Activation in the right hemisphere was generally correlated with the reaction time. They also point out an extensive network of brain regions associated with visual processing, motor control, and visual attention that is activated during the test, while some areas, such as posterior cingulate gyrus are deactivated (Ogg et al., 2008).

Behavioral CPT Performance and Brain Correlates in Schizophrenia Patients

In the CPT test the respondent has to decide whether to respond or not, as well as the maintaining arousal and self-monitoring of behavior (self-control) are needed. Schizophrenia patients are prone to attention/vigilance impairments and self-monitoring dysfunction (e.g., Stirling et al., 1998). A large multi-site study of the Consortium on the Genetics of Schizophrenia (COGS) showed that schizophrenia patients performed poorly compared to healthy subjects, even when controlled for differences in age, sex, education, and racial distribution (Gur et al., 2007). The studies focusing on neurocognitive deficits in schizophrenia and including CPT measures are consistent in their results, showing more commission and omission errors in schizophrenia patients (Earle-Boyer et al., 1991; Elvevag et al., 2000).

First study combining fMRI with CPT task in schizophrenia was performed by Volz et al. (Volz et al., 1999). Systematic review of scientific literature on fMRI studies using a sustained attention task was published by Sepede et al. (Sepede et al., 2014). The review included 11 studies of patients with schizophrenia, of which four studies used the CPT test paradigm: 2 studies used CPT-X (i.e., Eyler et al., 2004; Honey et al., 2005) and 2 studies CPT-IP (i.e., Volz et al., 1999; Salgado-Pineda et al., 2004). A recent imaging study applied an fMRI paradigm of the dual response AX-CPT test version (i.e., Lesh et al., 2013). Significant differences in activation patterns between patient and control groups were found in all studies selected by Sepede et al. (Sepede et al., 2014), even in case the patients performed comparable to the control group (e.g., Eyler et al., 2004). All of the mentioned studies support the finding of the attentional deficit in schizophrenia tested in CPT variants. This deficit was mainly related to hypoactivity in anterior and posterior cingulate cortex and in the right prefrontal cortex (Sepede et al., 2014). Thalamus activation results are inconsistent, reported to be either hyperactivated in SZ patients during CPT (Honey et al., 2005), or hypoactivated (Volz et al., 1999; Salgado-Pineda et al., 2004). Altered activation pattern was also reported in the thalamus. However, both thalamic hypoactivation and hyperactivation have been reported (for more details on imaging studies in schizophrenia see **Table 1**).

VERBAL FLUENCY (VFT)[3]

Behavioral Performance (Errors) and Brain Correlates of VFT in Healthy Individuals and in Lesion Studies

Verbal fluency tests (VFT) require the subject to produce words according to a specific rule and were designed to evaluate several aspects of verbal behavior including cognitive flexibility, switching response sets, self-regulating and self-monitoring (Lezak et al., 2012). Primarily, the VFT assesses higher functions of verbal organization and management (Bertola et al., 2014). The typical two verbal fluency tests defined by Lezak et al. (2012) and Laine (1988) are a category fluency and a letter fluency. In category or semantic fluency test, the subject is required to generate a list of words that are associated with their meaning (e.g. a list of animals or fruits). In letter or phonemic fluency, phonological clusters are made that are either words with the same initial letter or homonyms (e.g., fair, fare) (Lezak et al., 2012). Participants are not allowed to repeat the same word thus indirectly assessing their short-term memory, since they have to remember which words have already been said (Estes, 1974; Lezak et al., 2012; Fischer-Baum et al., 2016). Other significant cognitive functions that the VFT assesses are lexical-semantic knowledge and automatic retrieval (Hurks et al., 2010), controlled information processing (Hurks et al., 2006), sustained attention, strategic planning, searching and inhibition (Birn et al., 2010).

Qualitative errors encountered in a verbal fluency test include **breaking set errors (intrusions)** and **repetition (perseverations) errors**. The latter can be further divided (Galaverna et al., 2016) into simple repetitions, true perseverations and using the same stem in two words (e.g. "paint," "painter"). **True perseverations** occur in consecutive words whereas simple repetitions are made after a few seconds, possibly showing decreased searching and inhibition skills or even deficits in working memory (Azuma, 2004; Lezak et al., 2012; Fischer-Baum et al., 2016). **The intrusions or breaking set errors** refer to inappropriate responses with production of words with a different initial letter or different category than the assigned one (Lezak et al., 2012; Galaverna et al., 2016). Intrusions are associated with decreased inhibition and/or increased susceptibility to interference (Mahone et al., 2001).

Chertkow and Bub (1990) noted that generating words beginning from the same letter is not practiced as a skill in everyday life thus requiring strategic thinking whereas category

---

[3]Performance on the VFT is measured by the number of words generated within a set time limit, usually, 1 min, and normative data have been collected for different age groups, suggesting average ranges of words/min that should be expected from healthy subjects (Lezak et al., 2012).

TABLE 1 | summarizes the brain correlates and test performance in schizophrenia.

TASK	Study references	Task variant	Study sample	Imaging method	Reported brain areas	Task-related results and comments
TMT	Horacek et al., 2006	TMT A and B	SZ (N = 42), HC(N = 42)	(18) FDG positron emission tomography (PET), TMT A and B performance as covariates.	Higher metabolic activity in temporal, parietal, pre- and postcentral gyri, precuneus, limbic regions (anterior cingulate, uncus) and pons predicted better performance on TMT B (independent of the group).	TMT A and TMT B "time"—SZ (↓↓) vs. HC.
	Argyelan et al., 2013	TMT A	SZ (N = 18), BD (N = 19), HC(N = 32)	Resting state fMRI	Decreased functional connectivity of the left caudate nucleus, temporal occipital fusiform cortex/lingual gyrus, left thalamus predicts worse performance on TMT A in SZ patients (effect gradient: SZ>BD>HC)	TMT A "time"—SZ(↓) and BD vs. HC.
	Schuepbach et al., 2016	TMT A and B	FEP (N = 15), HC (N = 15)	Functional transcranial Doppler sonography	The blood flow velocity (BFV) measured in middle cerebral arteries during TMT B was significantly increased in patients in comparison to HC. In contrast to patients, BFV of HC subjects returned during TMT to low initial levels.	TMT A and TMT B "time"—SZ (↓↓) vs. HC, "number of errors" (no group differences).
	Liu et al., 2017	TMT A	Adolescent-onset SZ (AOS) (N = 48), HC (N = 31)	rsfMRI, voxel-mirrored homotopic connectivity (VMHC) and support vector machine (SVM) analyses	AOS group vs. HC exhibited decreased VMHC values in the following brain regions: fusiform gyrus, superior temporal gyrus (STG)/insula, precentral gyrus, and precuneus.	TMT A "time"—SZ (↓) vs. HC Decreased VMHC values in the superior temporal gyrus/insula correlated with slower TMT-A performance.
CPT	Volz et al., 1999	CPT-IP	SZ (N = 14), HC (N = 20)	fMRI	SZ patients exhibited decreased CBF in the right medial PFC, the right cingulate cortex and the left thalamus in comparison to HC.	CPT-IP performance—SZ vs. HC (no group effect) CPT-IP performance in male subjects was associated with activation pattern (poor performers with SZ showed more active brain regions than HC group and good performers with SZ. Poor performers with SZ vs. poor HC performers decreased dorsolateral PFC activation, increased in right temporal lobe).
	Eyler et al., 2004	CPT-X	Chronic SZ (N = 9), HC (N = 10)	fMRI	Activation of the frontal cortex was impaired in SZ compared to HC, specifically in the inferior frontal gyrus. These findings confirm the results of (Volz et al., 1999) and (Lesh et al., 2013), however, the localization of the prefrontal deficit differs.	CPT-X performance—SZ vs. HC (no group effect).
	Honey et al., 2005	CPT-X (addition of degraded stimuli to test cognitive dysmetria)	SZ with positive and negative symptoms (N = 11), SZ with predominant positive symptoms (N = 11), HC (N=12)	fMRI	SZ group showed abnormally high response to the CPT task in thalamus, left caudate nucleus and in frontal and temporal regions. Contrary, hypoactivation was observed in the middle frontal gyrus, putamen and angular gyrus.	CPT-X performance—SZ vs. HC: "target discrimination" (↓) in SZ, "reaction time" (no group effect) SZ subjects failed to elicit a task-related activation in response to the demands of the degraded stimuli. The pattern of task-related connectivity was disrupted in SZ.

(Continued)

TABLE 1 | Continued

TASK	Study references	Task variant	Study sample	Imaging method	Reported brain areas	Task-related results and comments
	Lesh et al., 2013	AX-CPT test version	FEP (N = 43), HC (N = 54)	fMRI	The dorsolateral PFC (DLPFC) activation in FEP was significantly associated with disorganization symptoms and performance during proactive cognitive control measured by CPT [AX-CPT]. SZ vs. HC: reduced recruitment of DLPFC and parietal cortex during CPT.	AX-CPT performance—FEP (↓↓) vs. HC.
VFT	Baaré et al., 1999	Category and letter VFT	SZ (N = 14), HC (N = 14)	Structural MRI, VBM	Lower semantic fluency scores in SZ patients correlated with smaller volumes of prefrontal gray matter bilaterally.	SZ vs. HC: letter VFT score (↓), semantic VFT score (↓↓) in SZ.
	Sanfilipo et al., 2002	Letter VFT	Male SZ medically stable (N = 62), HC (N = 27)	High-resolution MRI	Lower VFT scores in SZ correlated with smaller gray matter volumes in the prefrontal and temporal lobes bilaterally.	VFT score—SZ (↓↓) vs. HC Inverse correlations observed between cognitive abilities (psychomotor speed, cognitive flexibility and verbal fluency) and negative symptom.
	Weiss et al., 2004	Letter VFT	High-functioning SZ (N = 9), HC (N = 9)	fMRI	Bilateral activation inferior frontal gyrus area in a group of SZ patients, in contrast to the activation in left Broca's area seen in HC.	Letter VFT score—SZ vs. HC (no group effect)
	Boksman et al., 2005	Letter VFT	FEP (N = 10), HC (N = 10)	fMRI	HC group showed activations in the posterior parietal lobe, occipital lobe and cerebellum that were not detected in SZ patients.	Letter VFT "words"—FEP (↓) vs. HC.
	Fu et al., 2005	Letter VFT with 4 different letter sets (2x easy, 2x difficult conditions)	SZ, acute psychosis (N = 9), SZ in remission (N = 10), HC (N = 11)	fMRI	Increasing task demands (difficulty) led to greater anterior cingulate and right middle frontal activation in patients with active psychosis than in patients in remission. Decreased activation in SZ compared to HC in the anterior cingulate and the right prefrontal cortex (inferior and middle frontal cortices) independent of psychotic state and task demands.	VFT "number of errors"—SZ vs. HC (no group effect); in both groups, more errors observed during the difficult condition.
	Weiss et al., 2006	Letter VFT	Unmedicated SZ patients during an acute psychotic episode (N = 7)	fMRI	Bilateral activation of inferior frontal gyrus in SZ patients group was associated with the impaired verbal fluency performance.	Letter VFT score—SZ (↓) vs. HC. No evidence of decreased language-related activity in the left hemisphere of the SZ group. Gender differences in Processing strategies for phonemic verbal fluency tests were obtained.
	Takizawa et al., 2008	Letter VFT	SZ (N = 55), HC (N = 70)	52-channel near-infrared spectroscopy (oxy-hemoglobin concentration)	SZ patients' performance associated with slower and reduced increase in the prefrontal activation as compared to HC.	Letter VFT—SZ (↓↓) vs. HC.
	Bhojraj et al., 2009	Letter VFT and category VFT (20 s trials) from *Multilingual Aphasia Examination: Manual of instructions* (Benton and Hamsher, 1978)—total VFT scores	HC: young healthy adults- high genetic risk for SZ (N = 60), no-risk HC (N = 42)	MRI surface based volumetry (Freesurfer)	The high-risk subjects had verbal fluency deficits and decreased volumes of gray matter in the left pars triangularis, left supramarginal gyrus and right angular and Heschl's gyri.	High risk subjects vs. no-risk HC: "letter VFT (↓), category VFT (no group effect)". Left over right hemispheric asymmetry in the pars triangularis of the no-risk HC group (reversed in high risk HC group) suggested as a predictor for improved performance on VFT in both groups.

(Continued)

TABLE 1 | Continued

TASK	Study references	Task variant	Study sample	Imaging method	Reported brain areas	Task-related results and comments
	Lynall et al., 2010	Letter VFT (FAS version of the Controlled Oral Word Association Test)	SZ ($N = 12$), HC ($N = 15$)	rs-fMRI	In SZ, the reduced degree and clustering were locally significant in medial parietal, premotor, cingulate, and right orbitofrontal cortical nodes of functional networks. Functional connectivity and topological metrics correlated with each other and with behavioral performance on VFT.	Letter VFT—SZ vs. HC (no group effect) When using a DS score as covariate, functional brain networks resulted in the reduced clustering and small-worldness and reduced probability of high degree hubs in the SZ group.
	Meijer et al., 2011	Category VFT	Ultra-high risk (UHR) of psychosis subjects ($N = 37$)	Structural MRI, VBM	Lower VFT scores correlated with lower density in the gray matter of the right superior and middle temporal gyrus, right insula, and left anterior cingulate cortex.	Performance on the VFT may correspond to structural alterations in the brain. No difference in VFT performance between UHR subjects who developed a psychotic illness in the 2-year follow-up and subjects who did not transit to psychosis.
	Vandevelde et al., 2017	Letter VFT	SZ ($N = 15$), BD ($N = 14$), HC ($N = 20$)	fMRI	Study identified three activation clusters. In patients with schizophrenia there was reduced connectivity in a specific medio-prefronto-striato-thalamic network, unlike the bipolar patients and HC group. Suggesting a potential diagnostic significance of this network in SZ.	No group effect in VFT performance. Functional disturbances and different brain connectivity patterns may also be an etiologic factor of poor VFT performance.
DS	Minatogawa-Chang et al., 2009	DS forward and backward	FEP ($N = 88$), HC ($N = 86$)	Structural MRI, VBM	Correlations between the performance in the neuropsychological tests and the GM volume in the DLPFC as well as lateral parietal and superior temporal gyrus.	DS score—SZ vs. HC (group comparison not reported).
	Lynall et al., 2010	DS forward and backward	SZ ($N = 12$), HC ($N = 15$)	rs-fMRI, DS score as covariate The study measured aspects of both FC and functional network topology.	There were no significant associations between any connectivity or topological metrics and either forward or backward DS scores.	DS score—SZ vs. HC (group comparison not reported).
AVLT	Crespo-Facorro et al., 1999	RAVLT A/B recall—recall of both well-learned and novel word lists (A and B list of the Rey AVLT)	SZ ($N = 14$), HC ($N = 13$)	PET	Patients fail to activate cortical-cerebellar-thalamic-cortical circuitry during recall of both well-learned and novel word lists.	RAVLT "practiced and novel conditions"—SZ vs. HC (no group effect)
	Hazlett et al., 2000	CVLT based—errors related activity	Unmedicated SZ ($N = 20$), HC ($N = 32$)	fMRI	Serial-ordering strategy was associated with decreased activity in frontal cortex and increased activity in temporal cortex. Patients also exhibited hypofrontality (lower ratio of frontal to occipital activations) compared with healthy group.	CVLT "number of recalled words"—SZ vs. HC ($\downarrow\downarrow$). More severe hypofrontality associated with increased perseveration errors in SZ.
	Hofer et al., 2003	Verbal recognition	SZ in remission ($N = 10$), HC ($N = 10$)	fMRI	SZ group exhibited lower activation of right dorsolateral and anterior PFC, right anterior cingulate and left lateral temporal cortex during encoding; less activation in bilateral prefrontal and lateral temporal cortex during recognition.	AVLT "recognition task accuracy"—SZ vs. HC (no group effect).

(Continued)

TABLE 1 | Continued

TASK	Study references	Task variant	Study sample	Imaging method	Reported brain areas	Task-related results and comments
	Ragland et al., 2004	Verbal encoding	SZ (N = 14), HC (N = 15)	fMRI	SZ group showed impairment in activation measured by BOLD signal in bilateral prefrontal cortex during encoding of words, but increase activation in parahippocampal gyrus.	AVLT "recognition task accuracy"—SZ (↓↓) vs. HC Words recognition showed decreased activation of the left PFC.
	Allen et al., 2009	Deese-Roediger-McDermott false memory task	At-risk mental state (ARMS) subjects (N = 18) HC (N = 22)	fMRI	The high-risk group showed decreased activation in the medial temporal cortex and prefrontal regions, during both verbal encoding and recognition. These differences were associated with decreased recognition performance and the increased risk of psychosis.	ARMS vs. HC: "target items accuracy" (↓), lure words accuracy (ns.), "false recognition responses for novel words" (↓) Correct recognition relative to false alarms was associated with differential engagement of the hippocampus bilaterally in HC, this difference was absent in the high-risk group.
	Hurlemann et al., 2008	RAVLT	Early prodromal states (EPS) (N = 20), late prodromal states (N = 16), HC (N = 30)	Structural MRI, VBM	Reduction of hippocampal volumes in late but not early prodromal states correlates with poorer performance in RAVLT delayed recall.	The groups differed only in delayed recall LPS (↓) vs. EPS vs. HC

Arrows are applied in means of group performance evaluated (independent of the parameter measured in the task) and display significance level of performance (↓ - p<0.05; ↓↓ - p < 0.01). Abbreviations: FEP, first episode psychosis; SZ, schizophrenia; HC, healthy controls; SA, schizoaffective disorder; BD, bipolar disorder; EPS, early prodromal states; LPS, late prodromal states; DCS, transcranial direct current stimulation; fMRI, functional magnetic resonance imaging; PET, positron emission tomography; VBM, voxel-based morphometry; rsfMRI, resting state functional magnetic resonance imaging; CBF, cerebral blood flow; PFC, prefrontal corte.

fluency is based on "conceptual knowledge." This may be one of the reasons why category fluency scores are overall better than letter fluency, even in healthy control groups (Laws et al., 2010). Several studies have shown that increasing age and lower education level strongly correlate with poorer performance on category fluency, both in terms of volume of words generated and number of errors made (Mitrushina et al., 2005; Lezak et al., 2012). Importantly, Weiss et al. (2003) showed that, when controlling for performance differences, males and females show the same brain activation pattern during the verbal fluency task.

Shao et al. (2014) evaluated specific skills required for the VFT, including vocabulary knowledge, lexical access speed and executive control ability and how they correlate with category and letter fluency scores. Vocabulary knowledge and lexical access speed were shown to be more predictive of the category fluency performance compared to the letter fluency, while executive control ability did not have a significant effect on one variant over the other (Shao et al., 2014). The distinct aspects of cognitive performance that the VFT examines are further verified by studies showing that different brain regions are involved in letter and category fluency. In a meta-analysis of 31 studies, Henry and Crawford (2004) compared patients with focal cortical lesions to healthy controls. They showed that temporal lobe damage correlated with poorer performance on category fluency tasks whereas lesions in the frontal lobe negatively affected both category and letter fluency to the same extent (Henry and Crawford, 2004). Particularly left inferior frontal lobe was repeatedly demonstrated to be

active to a various extend in both VFT variants [meta-analysis and systematic reviews by Costafreda et al. (2006) and Wagner et al. (2014)]. A recent study highlighted the important role of the basal ganglia in both letter and category fluency. Chouiter et al. (2016) examined a group of 191 right-handed patients who had suffered a first, unilateral, focal lesion either in the left or right hemisphere. Results showed that letter and category fluency had certain identical regional associations in the left hemisphere, namely putamen, caudate nucleus, globus pallidus, superior and middle temporal gyri, angular gyrus, insula and parts of the supramarginal gyri, supporting the notion of a common word-producing mechanism (Chouiter et al., 2016). Additionally, letter fluency performance correlated with lesions in the rolandic operculum and the supramarginal gyrus unlike category fluency, which was preferentially affected by lesions in the posterior middle temporal gyrus and pallidum (Chouiter et al., 2016). Marien et al. (2001) proposed a significant role of the right cerebellum in retrieval and other non-motor language aspects, observing the significant linguistic deficits in patients with lesions in this area.

Behavioral Performance and Brain Correlates of VFT in Schizophrenia

Deficits in verbal fluency in schizophrenia patients are not a surprising finding but there are two competing theories on whether they should be attributed to a diminished access to the semantic store (Joyce et al., 1996) or to a disrupted

semantic store (McKay et al., 1996). In the former case, category and letter fluency scores should be equally affected whereas, according to the latter theory, category fluency performance would be lower than letter fluency, since it is more dependent on semantic memory. Henry and Crawford (2005) conducted a large meta-analysis on 84 studies, comparing verbal fluency scores between schizophrenia patients and healthy controls, and found that deficits in category fluency are more pronounced compared to letter fluency, thus supporting the concept of a compromised semantic store despite generally lower retrieval ability. They also proposed that category fluency can be a predictive test in estimating the probability of future psychosis development.

Volumetric differences in specific brain areas, such as gray matter volume in the prefrontal and temporal lobes, related to altered VFT performance in schizophrenia patients have been documented using both voxel-based morphometry (VBM) by Meijer et al. (2011) and high resolution MRI methods (Baaré et al., 1999; Sanfilipo et al., 2002). Functional disturbances and different brain connectivity patterns might also present an etiologic factor of poor VFT performance (for more details see **Table 1**). Several fMRI studies have indicated decreased activity in the right anterior cingulate cortex, prefrontal, inferior frontal and middle frontal lobe in patients with schizophrenia undergoing verbal fluency tasks (Boksman et al., 2005; Fu et al., 2005).

DIGIT SPAN (DS)[4]

Behavioral Performance and Brain Correlates of DS in Healthy and Lesion Cohorts

The Digit Span test (DS), which is part of the Wechsler batteries (the intelligence and memory scales), is widely used for measuring of the immediate and working memory. Beside standard performance scores, the errors that are possible to detect in DS, can be divided into two categories: **item errors** and **order errors**. In item errors there is a change in the length of the span. In order errors, there is a change in the order of the digits in the sequence but the length of the sequence remains the same. In the item error category, one can distinguish two **types of errors: omission and insertion**. These types of errors occur at the end of the span length (in DSF they occur at the beginning and in DSB in the end). They are also more common for longer sequences (Woods et al., 2011). **Omission** errors occur when a subject fails to repeat one or more digits of the sequence, while the rest of the span is in a correct order. Kaplan (1991) suggested that an omission in shorter span can be

---

[4]The DS test is composed of two different types, the Digit Span Forward (DSF), and Digit Span Backward (DSB), each testing different cognitive functions (Banken, 1985; Kaplan, 1991). In DSF test, the subject is requested to repeat a specific sequence of numbers, in the same order as presented. The DSB test follows the same principle and the main difference is that the subject is requested to repeat the sequence in a reverse order. DSB is an active procedure that requires effort, as it is composed by the encoding of the span, the manipulation, and reverse of the order and the recall of the correct digits, which makes it more demanding than DSF [(Banken, 1985; Black, 1986)].

a result of attention shifting, while an omission in longer span might signify a true memory deficit. An **insertion** error is the addition of an extra digit, which results in a longer span. In the order error category, there can be 3 different types of errors: sequencing, substitution, and repetition. These types of errors occur usually in the middle of the span and are more common for spans of shorter digits (Woods et al., 2011). **Sequence errors** present as a span with the correct length but a part of the span is in incorrect order (e.g., 1367 à 1637). **Substitution errors** occur when a subject replaces one digit for another, that could be part of the sequence or not (e.g., 23,578 and 23,478); here again, the span has the right length. **A repetition error** is the duplication of a digit, that appears in the same span (Woods et al., 2011).

The various MRI studies have helped to identify the brain activity associated with the performance of the DS. As demonstrated by Taki et al. (2011), the performance of DG positively correlated with the percentage of gray matter volume in the intracranial volume in the bilateral anterior temporal lobes. Another study that utilized voxel-based morphometry (VBM) and functional connectivity measures, confirmed the bilateral activation of anterior temporal lobes together with the left inferior frontal gyrus and the left Rolandic operculum, which constitute the critical areas in the auditory phonological loop of the verbal working memory (Goldman-Rakic, 1996). Along with the structural findings, DS scores were positively correlated with the resting state networks (rsN), namely the salience network (SN), that is, between the right anterior STG, the dorsal anterior cingulate cortex and the right fronto-insular cortex. It anti-correlated with the resting state functional connectivity (rsFC) within an anti-correlation network of the SN, between the right posterior superior temporal gyrus and the left posterior insula. Authors suggested such pattern of the activation reflected the neural organization of the phonological loop (Goldman-Rakic, 1996).

Another study demonstrated age-related and independent brain correlates with DS performance (Yang et al., 2015), specifically dorsal anterior cingulate gyrus showed distinctive roles in forward and backward span, whereas age dependent structures of the angular gyrus and sub-callosum were associated with DSF performance, and visual cortex and VLPFC were linked to DSB performance depending on age.

Lesion studies contribute to the understanding of the neural processes during DS performance challenging previously hypothesized neural targets. A study by Cave and Squire (Cave and Squire, 1992) showed in a sample of amnestic patients with hippocampal lesions and those with Korsakoff syndrome with a diencephalic damage that the DS scores of amnestic patients were performing close to controls, while the scores were significantly lower in the Korsakoff syndrome group. These findings support the view that the deficit in performance is independent of the hippocampal function. Another study challenged the role of the cerebellum in DS processing. In a single case study, a patient with a bilateral cerebellar ischemic lesion showed preserved DS performance (Chiricozzi et al., 2008).

Behavioral DS Performance and Brain Correlates in Schizophrenia Patients

Behavioral studies utilizing the digit span test have repeatedly shown a significantly impaired performance in individuals suffering from schizophrenia when compared to the healthy controls (Haenschel et al., 2009; Park and Gooding, 2014). In DSF the differences are controversial since some studies have shown that there are differences in span length between patients and healthy controls (Conklin et al., 2000; Galaverna et al., 2012) whereas others demonstrated similar performance in both groups (Moritz et al., 2001; Frydecka et al., 2016), suggesting the DSF test is not as demanding as DSB. On the other hand, the differences in DSB between patient and controls are quite significant. Several studies assessing working memory in individuals with schizophrenia using DS, have shown that not only the patients (Brébion et al., 2009) but also their first-degree relatives are able to remember shorter spans than the healthy controls (Conklin et al., 2000; Park and Gooding, 2014), which shows that working memory impairment is schizophrenia can have an endophenotypic character.

Studies reporting morphological or functional alterations associated with poor DS performance in schizophrenia are surprisingly rare (see **Table 1**). Minatogawa-Chang et al. (2009) reported significant correlations between the performance in DS task and the gray matter (GM) volume of DLPFC, parietal and temporal regions in first-episode psychosis (FEP) patients and healthy subjects. Interestingly, the middle frontal gyrus (BA46) GM volume was correlated only with the performance in FEP. On the other hand, the study by Lynall et al. (2010) failed to find any association between functional alterations in connectivity patterns measured during resting state and DS performance of schizophrenia patients. Studies reporting qualitative analyses of DS-related errors in schizophrenia and its association with functional or morphological changes are completely missing.

AUDITORY VERBAL LEARNING TASK (AVLT)[5]

Behavioral Performance (Errors) and Brain Correlates of AVLT in Healthy Individuals and in Lesion Studies

One of the most often applied learning and memory tests is the Rey Auditory-Verbal Learning Test (RAVLT) with a list containing 15 semantically unrelated words, contrary to the other AVLT variant, the California Verbal Learning Test (CVLT), which includes 16 semantically-related words (Mitrushina et al., 2005)[5]. The RAVLT method is very popular among clinicians for a good reason, as it allows to separate individual memory

processes that could be responsible for the identified disturbances of learning and memory.

The most commonly used measure in AVLT is the total number of correct responses (T_{1-5})[6] that informs us about the immediate recall score and, in terms of repeated trials, about the learning curve. Complete list of possible performance scores that can be calculated in RAVLT are well documented in Bezdicek et al. (2014). Here we will mention only the error-related qualitative approaches. The analysis of errors provides information about the memory processes and their integrity. Usually, only errors made during five consecutive learning trials (T1–T5) are reported (see e.g., Schmidt, 1996; Preiss et al., 2012). However, the quality of individual errors both during recall and recognition trials can also be recorded and analyzed. Two type of errors are usually detected: **Repetition errors or perseverations** are counted if the same correct word is listed more than once during one recall trial (recurring words), and are an important sign of impaired self-monitoring function (Lezak et al., 2012)[7]. **Intrusion errors** (confabulations or false productions) bring us more detailed information about the memory processes. Schnider et al. (1996) suggests that intrusions partially reflect the process of the effortful retrieval of memories despite the weak memory trace. Cunningham et al. (1997) suggested utilization of the so called "confabulation index" for quantification of confabulations in research studies, calculated as the proportion of novel recall intrusions to total responses. It is, however, important to distinguish provoked and spontaneous intrusions. Provoked intrusions from the list A to the interference list B or from the interference list B to post-interference recall of list A indicate sensitivity toward proactive interference and weakness of the context memory as suggested by Geffen et al. (1990). According to Barba et al. (2002), a weak memory trace may be a prerequisite for the occurrence of intrusions (promoted by interference at encoding). On the other hand, the extra-list intrusions (non-related or spontaneous intrusions) may be treated as a form of confabulation. Such intrusions may also reveal tendencies for semantic (category) or phonetic confusion of the original words (Mitrushina et al., 2005). Intrusion errors may thus serve as a measure of impaired executive functions applied in memory processes and should be analyzed in more detail, in order to prevent the vague interpretation produced by grouping all confabulations together (Cunningham et al., 1997). False recollections (increased number of intrusions) have been together with low recall performance previously described as a pattern typical for patients with focal frontal lobe lesions (Baldo et al., 2002) and dementia with prominent frontal lobe semiology (see Rouleau et al., 2001). Several studies also reported that false positives in recognition tests and intrusions on free recall trials are increased in confabulating patients (Bigler et al., 1989; DeLuca, 1993; Fischer et al., 1995; Cunningham et al., 1997). On the other hand, Nahum et al. (2012) shows that intrusions in memory tests have

---

[5]The standard administration of RAVLT includes five successive presentations (T1–T5) of the 15 words list (A) followed by free recall on each learning trial, (T6) presentation and recall of the "interference" list (B), a post-interference recall trial of the original list (A); and finally, the delayed recall (T7) and recognition (R) trials of the original list (A) with delay varying from 15 to 60 min (Mitrushina et al., 2005; Lezak et al., 2012; Preiss et al., 2012). Even that some studies do not apply the delayed recall or recognition trials, all parts of the test are clinically important.

[6]given over individual recall trials (T1+T2+T3+T4+T5).

[7]However, self-recognized repetitions, without an increased number of confabulations and intrusions, may be interpreted as an increased effort to recall as many words as possible (e.g., Bleecker et al., 2005).

no association with behaviorally spontaneous confabulations or disorientation.

Brain areas responsible for learning and memory as measured by verbal learning tasks, such as the RAVLT/CVLT, involve mainly frontal and temporal lobe areas, supported both by lesion and imaging studies (Savage et al., 2001; Baldo et al., 2002). Lesion studies, namely, Schouten et al. (2009) demonstrated that poor verbal memory performance as a result of the performance on both immediate and delayed recall and recognition in RAVLT[8] could be predicted by lesion characteristics. In their study patients with left hemispheric lesions, subcortical and large lesions performed poorly on the verbal memory measures. Medial temporal lobe (MTL) volume predicted the rate of learning in RAVLT in healthy volunteers as well (Fernaeus et al., 2013). Bilateral involvement of frontotemporal areas was also observed in studies that applied AVLT-based verbal memory fMRI paradigms. Johnson et al. (2001) provided evidence of the right frontal and left MTL involvement in verbal memory during CVLT task and documented a positive correlation between the activation of this network and task performance.

In terms of individual AVLT measures, the RAVLT first recall (Trial I) demonstrated the inferior parietal, middle frontal, and temporal activation (Wolk and Dickerson, 2011). Lezak et al. (2012) reported the involvement of MTL in last recall Trial V and hippocampal involvement during a delayed recall. More specifically, the head of the hippocampus is involved in verbal memory tasks (Hackert et al., 2002). Johnson et al. (2001) report the additional involvement of right anterior hippocampus. Recognition scores have been previously associated with the volume of perirhinal and entorhinal cortices Lezak et al. (2012) and right DLPFC activity, particularly in subjects with better memory abilities (Johnson et al., 2001). In healthy controls, the task of recalling the original list of 15-items after 24 h compared to resting baseline showed activation in frontal (left superior, and bilateral inferior and middle frontal gyrus) and parietal cortex (superior parietal gyrus bilaterally, right supramarginal gyrus) (Mensebach et al., 2009).

While the majority of studies focused on the exploration of cortical and hippocampal areas, other cortical and subcortical structures contribute to verbal memory too. Resting state functional connectivity (rsFC) study in a healthy population sample (Ystad et al., 2010) identified the correlations between CVLT measures and thalamic FC[9]. Another study on healthy subjects with memory complaints found RAVLT measures to be associated with glucose metabolism in posterior cingulate, precuneus, and orbitofrontal cortex (Brugnolo et al., 2014).

Behavioral Performance and Brain Correlates of AVLT in Schizophrenia

Verbal learning and memory deficits measured by AVLT tasks (related to frontotemporal dysfunction) have been repeatedly reported both in first episode schizophrenia subjects (FES)

(González-Blanch et al., 2007; Pérez-Iglesias et al., 2010; Rodriguez et al., 2015) and chronic schizophrenia patients (for review see Aleman et al., 1999; Boyer et al., 2007) and are considered as some of the main characteristics of cognitive deficits in schizophrenia (Keefe, 2008). One study assessing prodromal states of schizophrenia reported that reduction of hippocampal volumes in late but not early prodromal states correlates with poorer performance in RAVLT delayed recall (Hurlemann et al., 2008).

Even though the study sample of schizophrenia spectrum disorders often presents with a mixture of diagnoses, deficits in RAVLT performance may be a common denominator of the illness, as they are present in both paranoid and undifferentiated schizophrenia subtypes (Seltzer et al., 1997). The RAVLT performance is affected both in drug-free patients and patients on antipsychotic medication and is inversely correlated with negative symptoms (Manglam and Das, 2013), while no association with positive symptoms is observed. In contrary to CVLT, the RAVLT has been selected as a sensitive measure of outcome in schizophrenia (see Lepage et al., 2014) based on findings of several studies that failed to show differences in verbal memory between groups of patients with a different outcome (remitted vs. non-remitted) using CVLT. This could be due to semantically-related words that might help during the encoding process as was suggested by Lepage et al. (2014).

Despite the growing number of studies assessing verbal memory in schizophrenia using RAVLT, most of these studies only report total recall score (Trial 1–5; e.g., in Karilampi et al., 2007) and/or delayed recall performance in T7 (Pérez-Iglesias et al., 2010), while a minority of them report performance measured in particular trials (T1, T5, T6 for retention and recognition trial; e.g., Hurlemann et al., 2008). Specific characteristics of the RAVLT performance, such as errors, are often omitted in reported studies completely. Despite this lack of relevant RAVLT error-related literature, patients with schizophrenia show a higher total number of intrusions not affected by age, sex but correlating with patient IQ (Badcock et al., 2011). Our study performed in a group of FES patients showed an increased number of repetitions but not confabulations (intrusions) in comparison to a group of matched healthy volunteers (Rodriguez et al., 2015).

It has been suggested that errors in general or source memory deficits (repetitions of the correct answers and intra- and extra-list intrusions) underlie the positive symptoms of schizophrenia (Frith, 1995; Brébion et al., 2008, 2009). Some authors observed an association between the global number of extra- and intra-list intrusions and the positive symptoms score (Moritz et al., 2001) or thought disorder (Subotnik et al., 2006). In addition, the tendency to make false recognitions of non-target words may reflect a reality monitoring deficit associated with delusions and thought disorder as suggested by Ragland et al. (2003), and with hallucinations as reported in other studies (Brébion et al., 1998, 2005). In contrast, an inverse association was observed between the global number of intrusions and the negative symptom score (Heinrichs and Vaz, 2004) that might result from intensification of inhibitory processes that prevent intrusions. The higher number of extra-list intrusions during the free recall

---

[8]Both immediate and delayed recall and recognition in RAVLT.

[9]The study showed negative correlations between behavioral measures and the thalamus—functional connectivity of putamen and dorsomedial nucleus to thalamus and thalamus to caudate connectivity.

was negatively associated with certain negative symptoms, such as anhedonia, lack of spontaneity and emotional withdrawal (Brébion et al., 2002). Similar negative correlation with affective flattening was reported by Turetsky et al. (2002). As reported previously, mostly frontotemporal and subcortical networks are active during the performance in AVLT. These same networks are impaired in patients with schizophrenia. Several studies report functional or morphological changes in the medial temporal lobe and/or prefrontal areas related to observed disturbances in the AVLT performance (for more details see **Table 1**). While PFC activation was mostly reported only during encoding process (Hofer et al., 2003; Ragland et al., 2004), an error-related analysis in the study by Hazlett et al. (2000) showed that perseveration errors in schizophrenia are associated with hypoactivation of frontal areas.

GENERAL DISCUSSION

Theoretical Modeling of Generalized Cognitive Dysfunction in Schizophrenia by Classifying Errors and Underlying Brain Mechanisms

Standardized cognitive tests that are applied in schizophrenia research usually provide overall scores for individual tests, neglecting their qualitative characteristics. The test performance in healthy individuals, directly or indirectly supported by brain imaging data suggests the recruitment of the various cortical and subcortical brain structures, which also points to the compounded processing. Oftentimes, in order to explain the complex brain response, we find the explanation that proposes the involvement of various cognitive processes such as attention, memory executive functioning being themselves fairly complex. To date, there are no studies that would match the individual performances with the subject's brain activity, though this could be a desirable approach in clarifying the picture. Furthermore, the studies that are currently present are discrepant, each identifying novel brain structures involved during performance in particular cognitive task. Such a discrepancy in identified brain areas might be explained either by the heterogeneity of the tests' execution among patients or by the improvement of imaging techniques and statistical analysis. Whichever is true, a lack of a strong theoretical framework is obviously a disadvantage in the cognitive neuroscience of schizophrenia.

The attempt of modeling a cognitive dysfunction in schizophrenia proposed by Silverstein (2008) was based on the definition of specific and generalized cognitive dysfunction and was implemented in the large-scale schizophrenia research project CNTRICS. Since the neuropsychological tests are generally confounded by the multiple processes and a number of factors can impact cognitive functioning (fatigue, lack of motivation), in order to address specific deficits, Silverstein suggested: (1) to use a match task approach (application of two tasks that match on the variance and reliability) and (2) to apply process-specific task with the subsequent analysis of the changes across multiple conditions and multiple time points in order. Being complementary, the first approach would help to identify the specific deficits on the behavioral level, while the second—would help building up a mathematical model by application of the analysis of the covariance, principal component analysis, aggregation of scores into the cognitive subdomains, partially ordered sets and process-oriented strategies.

Indeed, the modeling approach has been widely used in cognitive psychology proposing the models for elementary cognitive processing including the reaction times (Townsend and Ashby, 1983) and more complex processing such as memory and reinforcement learning in healthy and clinical populations (Neufeld, 2007, 2015). Several studies aimed at mathematical modeling of the specific cognitive task on learning and memory [e.g., Continuous Presentation Task by Atkinson et al. (1967)], including the cognitive neuroimaging approaches in assessing the process of the decision making (Ahn et al., 2011; White et al., 2012).

Though Silverstein called for the specification of the deficits in schizophrenia, the mechanisms of the generalized deficits (possibly specific for some clusters of schizophrenia patients) seem to be overlooked. The approach based on the analysis of the similar errors that may occur in cognitive tests would allow to identify the common denominators of the generalized deficits. In the current review, we have gathered the characteristics of qualitative performance (errors) that can be detected in commonly used neuropsychological tests. From the summary provided, the similarities and differences in the errors across tests can be identified. For example, in the variety of tests, one can detect **perseverative errors or repetition errors** that are defined as the immediate inappropriate repetition of a prior response and are common for dorsolateral prefrontal cortex and basal ganglia dysfunction (Schindler et al., 1984; Hauser, 1999; Nys et al., 2006). However, Ramage et al. (1999) have shown that about four percent of the healthy cohort with both young and older subjects commit perseverations. Indeed, in children perseverations are normal and are attributed to the brain immaturity and lack of inhibitory mechanisms (Hauser, 1999). The development of the prefrontal cortex can reduce perseverations by supporting the strengthening of active representations in a competition between latent memory traces for previously relevant information and active memory traces for current information (Munakata et al., 2003). **Intrusion errors (form of confabulations)** refer to the inappropriate repetition of prior responses after intervening stimuli (Lorente-Rovira et al., 2011). Schindler et al. proposed that spontaneous confabulations (unprovoked) might be a result of a disconnection between orbitofrontal cortex (through the dorsomedial nucleus) with the amygdala (Schindler et al., 1984). Lesion studies suggest that confabulations are associated with damage in the right ventromedial frontal lobes, cingulate gyrus, cingulum, anterior hypothalamus, and head of the caudate nucleus (Moscovitch and Melo, 1997) and, similarly to perseverations, are detectable in healthy subjects (Burgess, 1996).

Further, **omission errors** correspond to the missing target; and **commission errors** imply the response to any stimulus other than the target as suggested by the instruction; those are typically detected in various GO/noGO task. In the study of Menon and Uddin (2010), the left and right insula and

adjoining inferior frontal cortex, right anterior cingulate, and left precuneus/posterior cingulate showed significantly greater activation during error processing (omission), compared to response inhibition (commission) and competition.

Lastly, sequential processing refers to the mental integration of the stimuli in a particular serial order. **Sequential errors** are presumably dependent on the cognitive domain (perceptual, motor). Thus, in the motor system, the underlying pattern of activation involves the primary motor and sensory areas, cerebellum, and basal ganglia (Ghilardi et al., 2009). Also, the sequential errors can be common in phonological processing (Whitaker, 1972). For example, Kuchinke et al. (2009) demonstrated that semantic processing of sequential relations additionally activated left medial and middle frontal gyrus, and left inferior frontal gyrus (Kuchinke et al., 2009). Therefore, each of the types of errors seem to have the unique pattern of the brain activity.

Importantly, besides generating performance errors, the human brain employs a meta-function aimed at monitoring the errors. This prefrontal monitoring system has been studied extensively, with its' center proposed in the anterior cingulate cortex (ACC). ACC is known to serve cognitive control functions enabling the brain to adapt the behavior in accordance to the changing task demands as well as the environmental circumstances (Botvinick et al., 2001).

Given the results of studies reported above, one can hypothesize that the brain encompasses the error detection system that prevents the errors occurrence and keeps monitoring the ongoing performance. Both the failure of error monitoring system and the errors described above can be detected across multiple tests indicating the generalized deficits that has been repeatedly reported in patients with schizophrenia (Goldstein and Shemansky, 1995). In other words, for instance, perseverations could be detected in several tests (not necessarily similar tests) that are predisposed to this type of errors. With respect to the brain activations that are associated with the specific errors, the pattern seems to lie in the central hubs of the cortex (DLPFC, OFC, ACC, PCC, precuneus) or within large scale networks (the description is below) with projections into subcortical structures (basal ganglia, amygdala). In line with error analysis, current evidence suggests an existence of a multiple error processing network in the brain, involving frontal and parietal regions and specifically ACC (Stevens et al., 2009). The structures are usually involved in the successful performance but the degree of activation fluctuates when the error occurs. From the experiment conducted, Stevens et al. also conclude that adults show a greater response amplitude in several error-related networks in comparison to adolescents suggesting that the normal maturation implements the greater responsiveness of the relevant brain structures to errors. Another study of Wierenga et al. (2015) suggested that the connectivity of the unimodal regions strengthens in childhood, while in adolescence the largest changes occur within and between frontal and parietal lobes, presumably indicating the greater flexibility of these regions. Indeed, it seems that the circuits require to sustain a certain level of activity in order to prevent the commission of the errors.

From the summary of the studies provided above, the cognitive test performance is not limited to error detection and "prevention" circuits. From our review of the task performance studies, it seems that many additional probably task/function-related structures are activated during the task performance. Moreover, the subsequent analysis in schizophrenia studies has revealed a diverse constellation of brain structures that are activated during the test performance or when correlating with cognitive performance (distinct from healthy individuals). How can this data discrepancy be interpreted? Firstly, assuming structural and functional brain alterations in schizophrenia, cognitive processing becomes more effortful creating the necessity of additional circuits to be involved. Secondly, in all scrutinized tests, patients exhibited a decreased activity in the medial temporal gyrus or superior temporal gyrus (Goldman et al., 2008). This finding is common in schizophrenia, and may not be related to the test performance *per se*, but rather indicate the group-specific and possibly symptoms related alterations. On the other hand, the alterations in cortical structures, parts of the associative cortex (temporal, parietal, occipital) could contribute to the specific cognitive deficits. The immense variability of the results (activation patterns) obtained from task-related performance studies therefore suggests more general error-related approach as more useful to model cognitive dysfunction in schizophrenia.

Optimizing the Research Strategies of Brain Network Analysis

MRI and related methods have been prolific in the identification of networks that may be associated with specific functions. Thus, in the resting state, the cerebral cortex produces consistent spatiotemporal patterns of activity (Damoiseaux et al., 2006). These spontaneously emerging fluctuations map the cortex in a similar way as they are produced during task performance (Deco et al., 2013) being in a "stand-by mode" and indicating the readiness of the system to respond to stimuli (Van Vreeswijk et al., 1994). Resting-state networks (RSNs) are slightly discrepant across different studies (Lowe et al., 1998; Cordes et al., 2001; Damoiseaux et al., 2006), though some studies classify the networks according to their functional role in cognitive processing. The large-scale networks include a salience network (SN), DMN and central executive network (CEN) each of which correspond to a specific functional role, though being functionally linked with each other. The SN involves the dorsal-anterior cingulate and anterior insula regions and is involved in the selection of salient external and interceptive signals (Sridharan et al., 2008). The DMN consists of midline structures, notably the medio-frontal cortex and posterior cingulate, being dominant in the resting state and deactivating during focused activity (van Buuren et al., 2010). The CEN includes the regions in the middle and inferior prefrontal and parietal cortices that are engaged in many higher-level cognitive tasks (Menon, 2011). In the task-related activity, these networks act consensually; CEN activates while being triggered by the externally oriented stimuli while DMN shows decreased activation. The SN causally influences

the activation of DMN/CEN by switching between these two networks (Nekovarova et al., 2014). Another classification of functional network was proposed by Power et al. who divided the networks into "processing" or "control" categories (Power et al., 2011). Processing-type networks are static and modular, and control networks are dynamic and adaptable to various tasks. Furthermore, frontal-parietal networks (FPN) that include the lateral prefrontal cortex and posterior parietal cortex presumably play a role in the top-down control (Dosenbach et al., 2008). One more classification of task-positive and task negative networks has been proposed (Fox et al., 2009). It is based on the assumption that the activity within the specific networks correlates positively or negatively based on analogous or opposite functional roles. Task-positive and negative networks are constituted of a set of regions that repeatedly increase-decrease in activity during attention-demanding stats (Fox et al., 2009).

Currently, different approaches for identifying patterns of coherent activity are used for the analysis of resting state networks (for a review see Cole et al., 2010). Functional and effective connectivity are concepts critical to this framework. Seed-to-voxel connectivity approach (Van Dijk et al., 2010) assists in identifying the specific connections that might be attributed to the type of errors in the cognitive tests. Given the fact that the specific type of errors can be detected across the tests (for example sequential errors in a Trail Making Test, Digit Span etc.), this connectivity approach might help to identify the associated brain connections.

A clustering-approach, based on single-subject independent component analysis (ICA) has been introduced by Esposito et al. (2005). The algorithm introduces a complex similarity measure by taking into account spatial and temporal characteristics for clustering. As temporal RSN patterns do not imply very diverse temporal characteristics this leads to unpredictable outcomes.

Functional connectivity fMRI explores correlations between time series from one region of interest (ROI) to another but cannot make inferences about influences between these regions. One can estimate causality of direction between activation in one brain node and another one with the methods based on effective connectivity. The most used means to estimate effective connectivity are Structure Equation modeling (SEM), Psychophysiological interactions (PPI) and Granger Causality (GC)—(can make inferences about linear states) (for review see Stephan and Friston, 2010). There are several differences between these approaches. The Granger Causality models the dependence among observed neural responses or patterns of activity (Friston et al., 2013). The Graph theory envisages the brain as a networked system composed of nodes and links. The various brain sites and anatomical tracts connecting them and their interrelations are encoded in the measures of the statistical dependencies. The networks architecture presumes the existence of the neural hubs, referring to the brain areas that are localized centrally and are densely connected with the other structures, together constituting a "rich club" (van den Heuvel and Sporns, 2013). The rich club networks are also referred to as large-scale networks, notably DMN, CEN and salience network. These networks hold long-range connections between the distant brain areas. On the other hand, the "small-world" networks consist of dense local clusters of connections between neighboring nodes and have a short path length between distant pair of nodes, at the same time being more specialized in function. Over the past years, several reports have consistently suggested that brain hubs and their rich club connections imply the efficient neural communication and integration, constituting "a central communication backbone that boosts the functional repertoire of the system" (van den Heuvel and Sporns, 2013). Schizophrenia patients exhibit the reduced connectivity between the rich club nodes of the brain as well as the small world nets. The same patterns are also present in siblings of patients and in their healthy offsprings (van den Heuvel and Sporns, 2013). Using the information on the networks from the previous step, the number and the properties of connections (centrality, assortativity, transitivity and path length etc.) within and between the defined networks using graph theoretical approach can be explored.

Patients with schizophrenia tend to have a less integrated functional brain connectivity (Lynall et al., 2010). However, the current status quo misses the aspect of the interactions between and within the cognitive functions and brain circuits. In the early review, Pantelis and Brewer (1996) provide an example of the study that tapped into specific errors in the performance. Thus, Owen et al. (1990) have decomposed the analysis of the set-shifting task (WCST) based on the failure in set-shifting or perseverations. This strategy has helped to identify the brain circuits associated with set-shifting (cortical-subcortical axis, basal ganglia) and perseveration errors (frontal), respectively. In this sense, they proposed a dichotomy between component-specific (cortical-subcortical networks that can be prompted by decomposing the complex cognitive functions) and network-specific (frontal-striatal-thalamic circuits coupled with the specific/solid pattern of function or behavior) brain functions. Twenty years later, Sheffield and Barch (2016) highlight the circuits that could serve as a reference point to study cognitive deficits in schizophrenia: (a) task-positive and task-negative functional brain (DMN, frontal-parietal, cingulo-opercular networks); (b) Cortico-Cerebellar-Thalamic-Cortical Circuit (CCTCC) to support main cognitive abilities (DMN, frontal-parietal, cingulo-opercular networks, subcortical networks and cerebellum) (c) the Go/NoGo pathway of reinforcement learning (encompassing the CCTCC networks+activation in the striatum). They point out the necessity to examine interactions between systems in schizophrenia, since the complexity and a range of dysfunctions are hardly due to single system impairments (Barch and Ceaser, 2012). Referring to the recent literature on connectomics (Park and Friston, 2013), it is less likely that only one part of the brain could be responsible for errors during performance. Rather, there is a probability of specific networks impairment. Although it is difficult to derive specific networks from studies above, we can see several areas of the brain that could serve as hubs for these networks.

Limitations

Several limitations of the study should be mentioned. Firstly, the temporal characteristics of the cognitive processing is not discussed in the review, though it can impact greatly the actual outcome of the brain analysis (Smith et al., 2012). Large

scale networks that encompass long-distance projections usually create a faster dynamic and are easily detected in contrast to the local networks (Shaposhnyk and Villa, 2012). However, the local processes presumably occur in the shorter time windows and might be problematic for detection with the current MRI technology. The interactions between the systems are also not mentioned in the current review though the experimental evidence is limited and does not allow to make any predictions (Wang et al., 2016; Dixon et al., 2018; Senden et al., 2018). Moreover, the application of simultaneous EEG-fMRI might be beneficial since to date no studies of this kind exist. In addition, the use of Granger causality methods could be problematic in fMRI as the hemodynamic response function is different between brain areas. This could be confounder in the temporal precedence of neuronal events as well as lower sampling rates or noise (Bajaj et al., 2016). While method is dependent on the selected model there is risk of spurious influence in eluding region, that drives the interactions in the model. The basic model of dynamic causal modeling (DCM) was enriched of modeling of neuronal fluctuations, called the stochastic modeling or spectral models. Thus, DCM could be used also for resting state fMRI data (Frässle et al., 2018). The debate is whether model involves all possible biological knowledge to model neuronal function. Some authors suggest that for example activity dependent plasticity or back-propagation is neglected in standard DCM (Daunizeau et al., 2011) But, Daunizeau also questioned if these specific "fine grained" mechanisms could be captured in BOLD signal. The other limitation of DCM is that the model is limited to maximum 10 regions, though new regression DCM method could possibly extend to whole-brain connectome analysis (Frässle et al., 2018).

Secondly, the selection of the proposed cognitive tasks was driven by the following arguments: (1) all the tasks are a part of the routine neuropsychological examination and are incorporated into the majority of the cognitive batteries used in the cognitive assessment of the patients with schizophrenia; (2) the tests do not explicitly tap into decision making process, requiring only simple manipulation with numbers, letters or words without making a choice (except for CPT test assessing also basic inhibition processes); (3) it was possible to dissect the tests based on the specific types of errors and to track common errors across the tests which served for the modeling purposes. Cognitive mechanisms or cognitive models of other cognitive tasks [for instance, Go/No-Go task (Yechiam et al., 2006)], Iowa Gambling Task (Fridberg et al., 2010), Wisconsin Card Sorting

test (Bishara et al., 2010) or Stroop task (Taylor et al., 2016) should be considered in the future analysis.

CONCLUSION

Investigation of interconnections between brain networks and cognitive functioning referring to cognitive deficits in schizophrenia on different levels (behavioral (cognitive performance) and physiological (brain networks) of disruptions is currently in progress though it requires a more rooted direction. By decomposing cognitive tests into more simple and accessible constructs and using additional qualitative characteristics of the performance, one can assort the related brain activity. Since the cognitive tests performances are often characterized by multiple errors, which can also be indirectly seen from a variety of brain activations, the possible approach could be to scrutinize the performance and to match the behavioral and neural patterns with the help of the recent mathematical modeling and connectivity tools. The error monitoring system should be also taken into account when investigating the complex brain-behavioral interactions in healthy subjects and in schizophrenia patients.

AUTHOR CONTRIBUTIONS

YZ has substantially contributed to the concept of the review, has made literature search and drafted the manuscripts. IF has contributed to the drafting of the manuscript and discussing and editing. BD, EB, IS, JM have written the parts of the manuscript. KŠ has written a part of the manuscript and edited the manuscript. MR and FŠ made critical points. JH has been involved into the conceptual discussion and provided the critical revisions of the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

REFERENCES

Ahn, A. A., Krawitz, A., Kim, W., Busmeyer, J. R., and Brown, J. W. (2011). A model-based fMRI Analysis with hierarchical parameter estimation. *J. Neurosci. Psychol. Econ.* 4, 95–110. doi: 10.1037/a0020684

Aleman, A., Hijman, R., de Haan, E. H., and Kahn, R. S. (1999). Memory impairment in schizophrenia: a meta-analysis. *Am. J. Psychiatry* 156, 1358–1366.

Allen, M. D., Owens, T. E., Fong, A. K., and Richards, D. R. (2011). A functional neuroimaging analysis of the Trail Making Test-B: implications for clinical application. *Behav. Neurol.* 24, 159–171. doi: 10.1155/2011/476893

Allen, P., Seal, M. L., Valli, I., Fusar-Poli, P., Perlini, C., Day, F., et al. (2009). Altered prefrontal and hippocampal function during verbal encoding and recognition in people with prodromal symptoms of psychosis. *Schizophr. Bull.* 37, 746–756. doi: 10.1093/schbul/sbp113

Andreasen, N. C., O'Leary, D. S., Cizadlo, T., Arndt, S., Rezai, K., Ponto, L. L., et al. (1996). Schizophrenia and cognitive dysmetria: a positron-emission tomography study of dysfunctional prefrontal-thalamic-cerebellar circuitry. *Proc. Natl. Acad. Sci. U.S.A.* 93, 9985–9990.

Antonova, E., Kumari, V., Morris, R., Halari, R., Anilkumar, A., Mehrotra, R., et al. (2005). The relationship of structural alterations to cognitive deficits in schizophrenia: a voxel-based morphometry study. *Biol. Psychiatry* 58, 457–467. doi: 10.1016/j.biopsych.2005.04.036

Argyelan, M., Ikuta, T., DeRosse, P., Braga, R. J., Burdick, K. E., John, M., et al. (2013). Resting-state fMRI connectivity impairment in schizophrenia and bipolar disorder. *Schizophr. Bull.* 40, 100–110. doi: 10.1093/schbul/sbt092

Atkinson, R. C., Brelsford, J. W., and Shiffrin, R. M. (1967). Multiprocess models for memory with applications to a continuous presentation task. *J. Math. Psychol.* 4, 277–300. doi: 10.1016/0022-2496(67)90053-3

Azuma, T. (2004). Working memory and perseveration in verbal fluency. *Neuropsychology* 18:69. doi: 10.1037/0894-4105.18.1.69

Baaré, W. F., Hulshoff Pol, H. E., Hijman, R., Willem, P. T., Viergever, M. A., and Kahn, R. S. (1999). Volumetric analysis of frontal lobe regions in schizophrenia: relation to cognitive function and symptomatology. *Biol. Psychiatry* 45, 1597–1605. doi: 10.1016/S0006-3223(98)00266-2

Badcock, J. C., Dragovic, M., Dawson, L., and Jones, R. (2011). Normative data for rey's auditory verbal learning test in individuals with schizophrenia. *Arch. Clin. Neuropsychol.* 26, 205–213.

Bajaj, S., Adhikari, B. M., Friston, K. J., and Dhamala, M. (2016). Bridging the gap: dynamic causal modeling and Granger causality analysis of resting state functional magnetic resonance imaging. *Brain Connect.* 6, 652–661. doi: 10.1089/brain.2016.0422

Baldo, J. V., Delis, D., Kramer, J., and Shimamura, A. P. (2002). Memory performance on the California Verbal Learning Test–II: findings from patients with focal frontal lesions. *J. Int. Neuropsychol. Soc.* 8, 539–546. doi: 10.1017/S135561770281428X

Banken, J. A. (1985). Clinical utility of considering Digits Forward and Digits Backward as separate components of the wechsler adult intelligence Scale-Revised. *J. Clin. Psychol.* 41, 686–691. doi: 10.1002/1097-4679(198509)41:5<686::AID-JCLP2270410517>3.0.CO;2-D

Barba, G. D., Mantovan, M. C., Traykov, L. T., Rieu, D., Laurent, A., Ermani, M., et al. (2002). The functional locus of intrusions: encoding or retrieval? *J. Clin. Exp. Neuropsychol.* 24, 633–641. doi: 10.1076/jcen.24.5.633.1008

Barch, D. M., and Ceaser, A. (2012). Cognition in schizophrenia: core psychological and neural mechanisms. *Trends Cogn. Sci.* 16, 27–34. doi: 10.1016/j.tics.2011.11.015

Benton, A. L., and Hamsher, K. (1978). *Multilingual Aphasia Examination: Manual of Instructions.* Iowa City, IA: University of Iowa.

Bertola, L., Mota, N. B., Copelli, M., Rivero, T., Diniz, B. S., Romano-Silva, M. A., et al. (2014). Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Front. Aging Neurosci.* 6:185. doi: 10.3389/fnagi.2014.00185

Bezdicek, O., Stepankova, H., Moták, L., Axelrod, B. N., Woodard, J. L., Preiss, M., et al. (2014). Czech version of Rey Auditory Verbal Learning test: normative data. *Aging Neuropsychol. Cogn.* 21, 693–721. doi: 10.1080/13825585.2013.865699

Bhojraj, T. S., Francis, A. N., Rajarethinam, R., Eack, S., Kulkarni, S., Prasad, K. M., et al. (2009). Verbal fluency deficits and altered lateralization of language brain areas in individuals genetically predisposed to schizophrenia. *Schizophr. Res.* 115, 202–208. doi: 10.1016/j.schres.2009.09.033

Bigler, E. D., Rosa, L., Schultz, F., Hall, S., and Harris, J. (1989). Rey-auditory verbal learning and rey-osterrieth complex figure design performance in Alzheimer's disease and closed head injury. *J. Clin. Psychol.* 45, 277–280.

Birn, R. M., Kenworthy, L., Case, L., Caravella, R., Jones, T. B., Bandettini, P. A., et al (2010). Neural systems supporting lexical search guided by letter and semantic category cues: a self-paced overt response fMRI study of verbal fluency. *Neuroimage* 49, 1099–1107. doi: 10.1016/j.neuroimage.2009.07.036

Bishara, A. J., Kruschke, J. K., Stout, J. C., Bechara, A., McCabe, D. P., and Busemeyer, J. R. (2010). Sequential learning models for the Wisconsin Card Sorting Task: assessing processes in substance dependent individuals. *J. Math. Psychol.* 54, 5–13. doi: 10.1016/j.jmp.2008.10.002

Black, F. W. (1986). Digit repetition in brain-damaged adults: clinical and theoretical implications. *J. Clin. Psychol.* 42, 770–782.

Bleecker, M. L., Ford, D. P., Lindgren, K. N., Hoese, V. M., Walsh, K. S., and Vaughan, C. G. (2005). Differential effects of lead exposure on components of verbal memory. *Occup. Environ. Med.* 62, 181–187. doi: 10.1136/oem.2003.011346

Boksman, K., Théberge, J., Williamson, P., Drost, D. J., Malla, A., Densmore, M., et al. (2005). A 4.0-T fMRI study of brain connectivity during word fluency in first-episode schizophrenia. *Schizophr. Res.* 75, 247–263. doi: 10.1016/j.schres.2004.09.025

Botvinick, M. M., Braver, T. S., Barch, D. M., and Carter, C. S., Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Boyer, P., Phillips, J. L., Rousseau, F. L., and Ilivitsky, S. (2007). Hippocampal abnormalities and memory deficits: new evidence of a strong pathophysiological link in schizophrenia. *Brain Res. Rev.* 54, 92–112. doi: 10.1016/j.brainresrev.2006.12.008

Brébion, G., David, A. S., Jones, H. M., and Pilowsky, L. S. (2009). Working memory span and motor and cognitive speed in schizophrenia. *Cogn. Behav. Neurol.* 22, 101–108. doi: 10.1097/WNN.0b013e3181a722a0

Brébion, G., David, A. S., Jones, H., and Pilowsky LS (2005). Hallucinations, negative symptoms, and response bias in a verbal recognition task in schizophrenia. *Neuropsychology* 19:612. doi: 10.1037/0894-4105.19.5.612

Brébion, G., Gorman, J. M., Amador, X., Malaspina, D., and Sharif, Z. (2002). Source monitoring impairments in schizophrenia: characterisation and associations with positive and negative symptomatology. *Psychiatry Res.* 112, 27–39. doi: 10.1016/S0165-1781(02)00187-7

Brébion, G., Ohlsen, R. I., Pilowsky, L. S., and David, A. S. (2008). Visual hallucinations in schizophrenia: confusion between imagination and perception. *Neuropsychology* 22:383. doi: 10.1037/0894-4105.22.3.383

Brébion, G., Smith, M. J., Amador, X., Malaspina, D., and Gorman, J. M. (1998). Word recognition, discrimination accuracy, and decision bias in schizophrenia: association with positive symptomatology and depressive symptomatology. *J. Nerv. Ment. Dis.* 186, 604–609. doi: 10.1097/00005053-199810000-00003

Brugnolo, A., Morbelli, S., Arnaldi, D., De Carli, F., Accardo, J., Bossert, I., et al. (2014). Metabolic correlates of Rey auditory verbal learning test in elderly subjects with memory complaints. *J. Alzheimers Dis.* 39, 103–113. doi: 10.3233/JAD-121684

Burgess, P. W. (1996). Confabulation and the control of recollection. *Memory* 4, 359–412. doi: 10.1080/096582196388906

Cave, C. B., and Squire, L. R. (1992). Intact verbal and nonverbal short-term memory following damage to the human hippocampus. *Hippocampus* 2, 151–163. doi: 10.1002/hipo.450020207

Chan, E., MacPherson, S. E., Robinson, G., Turner, M., Lecce, F., Shallice, T., et al. (2015). Limitations of the trail making test part-B in assessing frontal executive dysfunction. *J. Int. Neuropsychol. Soc.* 21, 169–174. doi: 10.1017/S135561771500003X

Chertkow, H., and Bub, D. (1990). Semantic memory loss in dementia of Alzheimer's type: what do various measures measure? *Brain* 113, 397–417.

Chiricozzi, F. R., Clausi, S., Molinari, M., and Leggio, M. G. (2008). Phonological short-term store impairment after cerebellar lesion: a single case study. *Neuropsychologia* 46, 1940–1953. doi: 10.1016/j.neuropsychologia.2008.01.024

Chouiter, L., Holmberg, J., Manuel, A. L., Colombo, F., Clarke, S., Annoni, J.-M., et al. (2016). Partly segregated cortico-subcortical pathways support phonologic and semantic verbal fluency: a lesion study. *Neuroscience* 329, 275–283. doi: 10.1016/j.neuroscience.2016.05.029

Cole, D. M., Smith, S. M., and Beckmann, C. F. (2010). Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Front. Syst. Neurosci.* 4:8. doi: 10.3389/fnsys.2010.00008

Collie, A., and Maruff, P. (2002). An analysis of systems of classifying mild cognitive impairment in older people. *Aust. N. Z. J. Psychiatry* 36, 133–140. doi: 10.1046/j.1440-1614.2002.00972.x

Conklin, H. M., Curtis, C. E., Katsanis, J., and Iacono, W. G. (2000). Verbal working memory impairment in schizophrenia patients and their first-degree relatives: evidence from the digit span task. *Am. J. Psychiatry* 157, 275–277. doi: 10.1176/appi.ajp.157.2.275

Conners, C. K. (2014). *Conners Continuous Performance Test 3rd Edition (Conners CPT 3) & Conners Continuous Auditory Test of Attention (Conners CATA): Technical Manual* (Multi-Health Systems Inc.).

Cordes, D., Haughton, V. M., Arfanakis, K., Carew, J. D., Turski, P. A., Moritz, C. H., et al. (2001). Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data. *Am. J. Neuroradiol.* 22, 1326–1333.

Costafreda, S. G., Fu, C. H., Lee, L., Everitt, B., Brammer, M. J., and David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: role of the left inferior frontal gyrus. *Hum. Brain Mapp.* 27, 799–810. doi: 10.1002/hbm.20221

Crespo-Facorro, B., Paradiso, S., Andreasen, N. C., O'Leary, D. S., Watkins, G. L., Boles Ponto, L. L., et al. (1999). Recalling word lists reveals "cognitive dysmetria" in schizophrenia: a positron emission tomography study. *Am. J. Psychiatry* 156, 386–392.

Cunningham, J. M., Pliskin, N. H., Cassisi, J. E., Tsang, B., and Rao, S. M. (1997). Relationship between confabulation and measures of memory and executive function. *J. Clin. Exp. Neuropsychol.* 19, 867–877. doi: 10.1080/01688639708403767

Damoiseaux, J. S., Rombouts, S., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., et al. (2006). Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13848–13853. doi: 10.1073/pnas.0601417103

Daunizeau, J., David, O., and Stephan, K. E. (2011). Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage* 58, 312–322. doi: 10.1016/j.neuroimage.2009.11.062

Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G. L., Hagmann, P., and Corbetta, M. (2013). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J. Neurosci.* 33, 11239–11252. doi: 10.1523/JNEUROSCI.1091-13.2013

DeLuca, J. (1993). Predicting neurobehavioral patterns following anterior communicating artery aneurysm. *Cortex* 29, 639–647. doi: 10.1016/S0010-9452(13)80287-0

Dixon, M. L., De La Vega, A., Mills, C., Andrews-Hanna, J., Spreng, N., Christoff, K., et al. (2018). Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention network. *Proc. Natl. Acad. Sci. U.S.A.* 115, 1598–1607. doi: 10.1073/pnas.1715766115

Dosenbach, N. U., Fair, D. A., Cohen, A. L., Schlaggar, B. L., and Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends Cogn. Sci.* 12, 99–105. doi: 10.1016/j.tics.2008.01.001

Earle-Boyer, E. A., Serper, M. R., Davidson, M., and Harvey, P. D. (1991). Continuous performance tests in schizophrenic patients: stimulus and medication effects on performance. *Psychiatry Res.* 37, 47–56. doi: 10.1016/0165-1781(91)90105-X

Elvevag, B., Weinberger, D. R., Suter, J. C., and Goldberg, T. E. (2000). Continuous performance test and schizophrenia: a test of stimulus-response compatibility, working memory, response readiness, or none of the above? *Am. J. Psychiatry* 157, 772–780. doi: 10.1176/appi.ajp.157.5.772

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., et al. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage* 25, 193–205. doi: 10.1016/j.neuroimage.2004.10.042

Estes, W. K. (1974). Learning theory and intelligence. *Am. Psychol.* 29:740. doi: 10.1037/h0037458

Eyler, L. T., Olsen, R. K., Jeste, D. V., and Brown, G. G. (2004). Abnormal brain response of chronic schizophrenia patients despite normal performance during a visual vigilance task. *Psychiatry Res. Neuroimaging* 130, 245–257. doi: 10.1016/j.pscychresns.2004.01.003

Fernaeus, S.-E., Julin, P., Almqvist, O., and Wahlund, L.-O. (2013). Medial temporal lobe volume predicts rate of learning in Rey-AVLT. *Adv. Alzheimers Dis.* 2:7. doi: 10.4236/aad.2013.21002

Fischer, R. S., Alexander, M. P., D'esposito M, and Otto, R. (1995). Neuropsychological and neuroanatomical correlates of confabulation. *J. Clin. Exp. Neuropsychol.* 17, 20–28.

Fischer-Baum, S., Miozzo, M., Laiacona, M., and Capitani, E. (2016). *Perseveration during verbal fluency in traumatic brain injury reflects impairments in working memory.*

Fornito, A., Yoon, J., Zalesky, A., Bullmore, E. T., and Carter, C. S. (2011). General and specific functional connectivity disturbances in first-episode schizophrenia during cognitive control performance. *Biol. Psychiatry* 70, 64–72. doi: 10.1016/j.biopsych.2011.02.019

Fox, M. D., Zhang, D., Snyder, A. Z., and Raichle, M. E. (2009). The global signal and observed anticorrelated resting state brain networks. *J. Neurophysiol.* 101, 3270–3283. doi: 10.1152/jn.90777.2008

Frässle, S., Yao, Y., Schöbi, D., Aponte, E. A., Heinzle, J., and Stephan, K. E. (2018). Generative models for clinical applications in computational psychiatry. *Wiley Interdiscip. Rev. Cogn. Sci.* 9:e1460. doi: 10.1002/wcs.1460

Fridberg, D. J., Queller, S., Ahn, W.-Y., Kim, W., Bishara, A. J., Busemeyer, J. R., et al. (2010). Cognitive mechanisms underlying risky decision-making in chronic cannabis users. *J. Math. Psychol.* 54, 28–38. doi: 10.1016/j.jmp.2009.10.002

Friston, K., Moran, R., and Seth, A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* 23, 172–178. doi: 10.1016/j.conb.2012.11.010

Frith, C. D. (1995). The cognitive abnormalities underlying the symptomatology and the disability of patients with schizophrenia. *Int. Clin. Psychopharmacol.* 10, 87–98.

Frydecka, D., Beszłej, J. A., Gościmski, P., Kiejna, A., and Misiak, B. (2016). Profiling cognitive impairment in treatment-resistant schizophrenia patients. *Psychiatry Res.* 235, 133–138. doi: 10.1016/j.psychres.2015.11.028

Fu, C. H., Suckling, J., Williams, S. C., Andrew, C. M., Vythelingum, G. N., and McGuire PK (2005). Effects of psychotic state and task demand on prefrontal function in schizophrenia: an fMRI study of overt verbal fluency. *Am. J. Psychiatry* 162, 485–494. doi: 10.1176/appi.ajp.162.3.485

Galaverna, F. S., Morra, C. A., and Bueno, A. M. (2012). Attention in patients with chronic schizophrenia: deficit in inhibitory control and positive symptoms. *Eur. J. Psychiatry* 26, 185–195. doi: 10.4321/S0213-61632012000300005

Galaverna, F., Bueno, A. M., Morra, C. A., Roca, M., and Torralva, T. (2016). Analysis of errors in verbal fluency tasks in patients with chronic schizophrenia. *Eur. J. Psychiatry* 30, 305–320.

Geffen, G., Moar, K. J., O'hanlon, A. P., Clark, C. R., and Geffen, L. B. (1990). Performance measures of 16–to 86-year-old males and females on the auditory verbal learning test. *Clin. Neuropsychol.* 4, 45–63.

Ghilardi, M. F., Moisello, C., Silvestri, G., Chez, C., and Krakauer, J. W. (2009). Learning of a sequential motor skill comprises explicit and implicit components that consolidate differently. *J. Neurophysiol.* 101, 2218–2229. doi: 10.1152/jn.01138.2007

Gilbert, E., Mérette, C., Jomphe, V., Emond, C., Rouleau, N., Bouchard, R. H., et al. (2014). Cluster analysis of cognitive deficits may mark heterogeneity in schizophrenia in terms of outcome and response to treatment. *Eur. Arch. Psychiatry Clin. Neurosci.* 264, 333–334. doi: 10.1007/s00406-013-0463-7

Golden, C. J., Espe-Pfeifer, P., and Wachsler-Felder, J. (2000). *Neuropsychological Interpretation of Objective Psychological Tests* (Fort Lauderdale, FL: Kluwer Academic Publishers; Nova Southeastern University).

Goldman, A. L., Pazewas, L., Mattay, V. S., Fischl, B., Verchiniski, B. A., Zoitick, B., et al (2008). Heritability of brain morphology related to schizophrenia: a large-scale automated magnetic resonance imaging segmentation study. *Biol. Psychiatry* 63, 475–483. doi: 10.1016/j.biopsych.2007.06.006

Goldman-Rakic, P. S. (1996). Memory: recording experience in cells and circuits: diversity in memory research. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13435–13437. doi: 10.1073/pnas.93.24.13435

Goldstein, G., Allen, D. N., and Seaton, B. E. (1998). A comparison of clustering solutions for cognitive heterogeneity in schizophrenia. *J. Int. Neuropsychol. Soc.* 4, 353–362.

Goldstein, G., and Shemansky, W. J. (1995). Influences of cognitive heterogeneity in schizophrenia. *Schizophr. Res.* 18, 56–69. doi: 10.1016/0920-9964(95)00040-2

González-Blanch, C., Crespo-Facorro, B., Álvarez-Jiménez, M., Rodríguez-Sánchez, J. M., Pelayo-Terán, J. M., Pérez-Iglesias, R., et al. (2007). Cognitive dimensions in first-episode schizophrenia spectrum disorders. *J. Psychiatr. Res.* 41, 968–977. doi: 10.1016/j.jpsychires.2006.08.009

Green, M. F., and Harvey, P. D. (2014). Cognition in schizophrenia: past, present, and future. *Schizophr. Res. Cogn.* 1, e1–e9. doi: 10.1016/j.scog.2014.02.001

Gur, R. E., Calkins, M. E., Gur, R. C., Horan, W. P., Nuechterlein, K. H., Seidman, L. J., et al. (2007). The consortium on the genetics of schizophrenia: neurocognitive endophenotypes. *Schizophr. Bull.* 33, 49–68. doi: 10.1093/schbul/sbl055

Hackert, V. H., den Heijer, T., Oudkerk, M., Koudstaal, P. J., Hofman, A., and Breteler MMB (2002). Hippocampal head size associated with verbal memory performance in nondemented elderly. *Neuroimage* 17, 1365–1372. doi: 10.1006/nimg.2002.1248

Haenschel, C., Bittner, R. A., Waltz, J., Haertling, F., Wibral, M., Singer, W., et al. (2009). Cortical oscillatory activity is critical for working memory as revealed by deficits in early-onset schizophrenia. *J. Neurosci.* 29, 9481–9489. doi: 10.1523/JNEUROSCI.1428-09.2009

Hagen, K., Ehlis, A. C., Haeussinger, F. B., Heinzel, S., Dresler, T., Mueller, L. D., et al. (2014). Activation during the Trail Making Test measured with functional near-infrared spectroscopy in healthy elderly subjects. *Neuroimage* 85, 583–591. doi: 10.1016/j.neuroimage.2013.09.014

Halperin, J. M., Wolf, L., Greenblatt, E. R., and Young, G. (1991). Subtype analysis of commission errors on the continuous performance test in children. *Dev. Neuropsychol.* 7, 207–217. doi: 10.1080/87565649109540488

Harvey, P. (1997). *Cognitive Function in Patients With Schizophrenia*. Report from a roundtable meeting of experts in cognitive dysfunction in schizophrenia, Budapest.

Hauser, M. D. (1999). Perseveration, inhibition and the prefrontal cortex: a new look. *Curr. Opin. Neurobiol.* 9, 214–222. doi: 10.1016/S0959-4388(99)80030-0

Hazlett, E. A., Buchsbaum, M. S., Jeu, L. A., Nenadic, I., Fleischman, M. B., Shihabuddin, L., et al. (2000). Hypofrontality in unmedicated schizophrenia patients studied with PET during performance of a serial verbal learning task. *Schizophr. Res.* 43, 33–46. doi: 10.1016/S0920-9964(99)00178-4

Heinrichs, R. W., and Vaz, S. M. (2004). Verbal memory errors and symptoms in schizophrenia. *Cogn. Behav. Neurol.* 17, 98–101. doi: 10.1097/01.wnn.0000116252.78804.73

Henry, J. D., and Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology* 18, 284–295. doi: 10.1037/0894-4105.18.2.2842

Henry, J. D., and Crawford, J. (2005). A meta-analytic review of verbal fluency deficits in schizophrenia relative to other neurocognitive deficits. *Cogn. Neuropsychiatry* 10, 1–33. doi: 10.1080/13546800344000309

Hofer, A., Weiss, E. M., Golaszewski, S. M., Siedentopf, C. M., Brinkhoff, C., Kremser, C., et al. (2003). An FMRI study of episodic encoding and recognition of words in patients with schizophrenia in remission. *Am. J. Psychiatry* 160, 911–918. doi: 10.1176/appi.ajp.160.5.911

Honey, G. D., Pomarol-Clotet, E., Corlett, P. R., Honey, R. A., McKenna, P. J., Bullmore, E. T., et al. (2005). Functional dysconnectivity in schizophrenia associated with attentional modulation of motor function. *Brain* 128, 2597–2611. doi: 10.1093/brain/awh632

Horacek, J., Dockery, C., Kopecek, M., Spaniel, F., Novak, T., Tislerova, B., et al. (2006). Regional brain metabolism as the predictor of performance on the Trail Making Test in schizophrenia. A 18FDG PET covariation study. *Neuroendocrinol. Lett.* 27, 587–594.

Hurks, P. P., Vles, J. S. H., Hendriksen, J. G. M., Kalff, A. C., Feron, F. J. M., Kroes, M., et al. (2006). Semantic category fluency versus initial letter fluency over 60 seconds as a measure of automatic and controlled processing in healthy school-aged children. *J. Clin. Exp. Neuropsychol.* 28, 684–695. doi: 10.1080/13803390590954191

Hurks, P. P., Schrans, D., Meijs, C., Wassenberg, R., Feron, F. J. M., and Jolles, J. (2010). Developmental changes in semantic verbal fluency: analyses of word productivity as a function of time, clustering, and switching. *Child Neuropsychol.* 16, 366–387. doi: 10.1080/09297041003671184

Hurlemann, R., Jessen, F., Wagner, M., Frommann, I., Ruhrmann, S., Brockhaus, A., et al. (2008). Interrelated neuropsychological and anatomical evidence of hippocampal pathology in the at-risk mental state. *Psychol. Med.* 38, 843–851. doi: 10.1017/S0033291708003279

Jacobson, S. C., Blanchard, M., Connolly, C. C., Cannon, M., and Garavan, H. (2011). An fMRI investigation of a novel analogue to the Trail-Making Test. *Brain Cogn.* 77, 60–70. doi: 10.1016/j.bandc.2011.06.001

James, G. A., Kearney-Ramos, T. E., Young, J. A., Kilts, C. D., Gess, J. L., and Fausett, J. S. (2016). Functional independence in resting-state connectivity facilitates higher-order cognition. *Brain Cogn.* 105, 78–87. doi: 10.1016/j.bandc.2016.03.008

Johnson, S. C., Saykin, A. J., Flashman, L. A., McAllister, T. W., and Sparling, M. B. (2001). Brain activation on fMRI and verbal memory ability: functional neuroanatomic correlates of CVLT performance. *J. Int. Neuropsychol. Soc.* 7, 55–62. doi: 10.1017/S135561770171106X

Joyce, E. M., Collinson, S. L., and Crichton, P. (1996). Verbal Fluency in Schizophrenia. *Psychol. Med.* 26, 41–52. doi: 10.1017/S0033291700033705

Kaplan, E. (1991). *WAIS-R as a Neuropsychological Instrument (WAIS-R NI) (Psychological Corporation)* (San Antonio, TX).

Karilampi, U., Helldin, L., Hjärthag, F., Norlander, T., and Archer, T. (2007). Verbal learning in schizopsychotic outpatients and healthy volunteers as a function of cognitive performance levels. *Arch. Clin. Neuropsychol.* 22, 161–174. doi: 10.1016/j.acn.2006.12.003

Katz, K. S., Dubowitz, L. M., Henderson, S., Jongmans, M., Kay, G. G., Nolte, C. A., et al. (1996). Effect of cerebral lesions on continuous performance test responses of school age children born prematurely. *J. Pediatric Psychol.* 21, 841–855. doi: 10.1093/jpepsy/21.6.841

Keefe, R. S. (2008). Should cognitive impairment be included in the diagnostic criteria for schizophrenia? *World Psychiatry* 7, 22–28. doi: 10.1002/j.2051-5545.2008.tb00142.x

Kopp, B., Rösser, N., Tabeling, S., Stürenburg, H. J., de Haan, B., Karnath, H.-O., et al. (2015). Errors on the trail making test are associated with right hemispheric frontal lobe damage in stroke patients. *Behav. Neurol.* 2015:309235. doi: 10.1155/2015/309235

Kraepelin, E. (1919). *Dementia Praecox and Paraphrenia*. Chicago, IL: E and S Livingstone.

Kubo, M., Shoshi, C., Kitawaki, T., Takemoto, R., Kinugasa, K., Yoshida, H., et al. (2008). Increase in prefrontal cortex blood flow during the computer version trail making test. *Neuropsychobiology* 58, 200–210. doi: 10.1159/000201717

Kuchinke, L., Meer, E. V., and Krueger, F. (2009). Differences in processing of taxonomic and sequential relations in semantic memory: an fMRI investigation. *Brain Cogn.* 69, 245–251. doi: 10.1016/j.bandc.2008.07.014

Laine, M. (1988). Correlates of word fluency performance. *Stud. Lang.* 12, 43–61.

Laws, K. R., Duncan, A., and Gale, T. M. (2010). 'Normal'semantic–phonemic fluency discrepancy in Alzheimer's disease? A meta-analytic study. *Cortex* 46, 595–601. doi: 10.1016/j.cortex.2009.04.009

Lepage, M., Bodnar, M., and Bowie, C. R. (2014). Neurocognition: clinical and functional outcomes in schizophrenia. *Can. J. Psychiatry* 59, 5–12. doi: 10.1177/070674371405900103

Lesh, T. A., Westphal, A. J., Niendam, T. A., Yoon, J. H., Minzenberg, M. J., Ragland, J. D., et al. (2013). Proactive and reactive cognitive control and dorsolateral prefrontal cortex dysfunction in first episode schizophrenia. *Neuroimage Clin.* 2, 590–599. doi: 10.1016/j.nicl.2013.04.010

Levin, H. S., Peters, B. H., Kalisky, Z., High, W. M. Jr., von Laufen, A., Eisenberg, H., et al (1986). Effects of oral physostigmine and lecithin on memory and attention in closed head-injured patients. *Cent. Nerv. Syst. Trauma* 3, 333–342. doi: 10.1089/cns.1986.3.333

Lezak, M. D., Howieson, D. B., Bigler, E. D., and Tranel, D. (2012). *Neuropsychological Assessment*. New York, NY: Oxford University Press.

Liu, Y., Zhang, Y., Lv, L., Hu, F., Wu, R., Zhao, J., et al. (2017).Decreased resting-state interhemispheric functional connectivity correlated with neurocognitive deficits in drug-naive first-episode adolescent-onset schizophrenia. *Int. J. Neuropsychopharmacol.* 21, 33–41. doi: 10.1093/ijnp/pyx095

Lorente-Rovira, E., Berrios, G., McKenna, P., Moro-Ipola, M., and Villagran-Moreno, J. M. (2011). Confabulaciones (I): Concepto, clasificación y neuropatología. *Actas EspPsiquiatr* 39, 251–259.

Lowe, M. J., Mock, B. J., and Sorenson, J. A. (1998). Functional connectivity in single and multislice echoplanar imaging using resting-state fluctuations. *Neuroimage* 7, 119–132. doi: 10.1006/nimg.1997.0315

Lynall, M. E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

Mahone, E. M., Koth, C. W., Cutting, L., Singer, H. S., and Denckla, M. B. (2001). Executive function in fluency and recall measures among children with Tourette syndrome or ADHD. *J. Int. Neuropsychol. Soc.* 7, 102–111. doi: 10.1017/S1355617701711101

Mahurin, R. K., Velligan, D. I., Hazleton, B., Mark Davis, J., Eckert, S., and Miller, A. L. (2006). Trail making test errors and executive function in schizophrenia and depression. *Clin. Neuropsychol.* 20, 271–288. doi: 10.1080/13854040590947498

Manglam, M. K., and Das, A. (2013). Verbal learning and memory and psychopathology in schizophrenia. *Asian J. Psychiatry* 6, 417–420. doi: 10.1016/j.ajp.2013.05.009

Marien, P., Engelborghs, S., Fabbro, F., and De Deyn, P. P. (2001). The lateralized linguistic cerebellum: a review and a new hypothesis. *Brain Lang.* 79, 580–600. doi: 10.1006/brln.2001.2569

McKay, A. P., McKenna, P. J., Bentham, P., Mortimer, A. M., Holbery, A., and Hodges, J. R. (1996). Semantic memory is impaired in schizophrenia. *Biol. Psychiatry* 39, 929–937. doi: 10.1016/0006-3223(95)00250-2

Meijer, J. H., Schmitz, N., Nieman, D. H., Becker, H. E., van Amelsvoort, T. A., Dingemans, P. M., et al. (2011). Semantic fluency deficits and reduced grey matter before transition to psychosis: a voxelwise correlational analysis. *Psychiatry Res. Neuroimaging* 194, 1–6. doi: 10.1016/j.pscychresns.2011.01.004

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* 15, 483–506. doi: 10.1016/j.tics.2011.08.003

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Mensebach, C., Beblo, T., Driessen, M., Wingenfeld, K., Mertens, M., Rullkoetter, N., et al. (2009). Neural correlates of episodic and semantic memory retrieval in borderline personality disorder: an fMRI study. *Psychiatry Res. Neuroimaging* 171, 94–105. doi: 10.1016/j.pscychresns.2008.02.006

Mesholam-Gately, R. I., Giuliano, A. J., Goff, K. P., Faraone, S. V., and Seidman, L. J. (2009). Neurocognition in first-episode schizophrenia: a meta-analytic review. *Neuropsychology* 23, 315–336. doi: 10.1037/a0014708

Minatogawa-Chang, T. M., Schaufelberger, M. S., Ayres, A. M., Duran, F. L., Gutt, E. K., Murray, R. M., et al. (2009). Cognitive performance is related to cortical grey matter volumes in early stages of schizophrenia: a population-based study of first-episode psychosis. *Schizophr. Res.* 113, 200–209. doi: 10.1016/j.schres.2009.06.020

Mitrushina, M., Boone, K. B., Razani, J., and D'Elia, L. F. (2005). *Handbook of Normative Data for Neuropsychological Assessment.* New York, NY: Oxford University Press.

Moll, J., Oliveira-Souza, R., de, Moll, F. T., Bramati, I. E., and Andreiuolo, P. A. (2002). The cerebral correlates of set-shifting: an fMRI study of the trail making test. *Arq. Neuropsiquiatr.* 60, 900–905. doi: 10.1590/S0004-282X2002000600002

Moritz, S., Lambert, M., Andresen, B., Böthern, A., Naber, D., and Krausz, M. (2001). Subjective cognitive dysfunction in first-episode and chronic schizophrenic patients. *Compr. Psychiatry* 42, 213–216. doi: 10.1053/comp.2001.23144

Moscovitch, M., and Melo, B. (1997). Strategic retrieval and the frontal lobes: evidence from confabulation and amnesia. *Neuropsychologia* 35, 1017–1034. doi: 10.1016/S0028-3932(97)00028-6

Munakata, Y., Morton, J. B., and Stedron, J. M. (2003). "The role of prefrontal cortex in perseveration: developmental and computational explorations," in *Studies in Developmental Psychology. Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks,* ed P. T. Quinlan (New York, NY: Psychology Press), 83–114.

Nahum, L., Bouzerda-Wahlen, A., Guggisberg, A., Ptak, R., and Schnider, A. (2012). Forms of confabulation: dissociations and associations. *Neuropsychologia* 50, 2524–2534. doi: 10.1016/j.neuropsychologia.2012.06.026

Nekovarova, T., Yamamotova, A., Vales, K., Stuchlik, A., Fricova, J., and Rokyta, R. (2014). Common mechanisms of pain and depression: are antidepressants also analgesics? *Front. Behav. Neurosci.* 8:99. doi: 10.3389/fnbeh.2014.00099

Neufeld, R. W. J. (2007). "Composition and uses of formal clinical cognitive science," in *Modeling Complex Systems: Nebraska Symposium on Motivation,* Vol. 52, eds B. Shuart, W. Spaulding, and J. Poland (Lincoln, NE: University of Nebraska Press).

Neufeld, R. W. J. (2015). "Mathematical modeling applications in clinical psychology," in *Oxford Handbook of Computational and Mathematical Psychology,* eds J. R. Busemeyer, J. T. Townsend, Z. Wang, and A. Eidels (Oxford: Oxford University Press), 341–368.

Nys, G., M., van Zandvoort, M. J. E., van der Worp, H. B., and Kappelle, L. J., de Han, E. H. F. (2006). Neuropsychological and neuroanatomical correlates of perseverative responses in subacute stroke. *Brain* 129, 2148–2157. doi: 10.1093/brain/awl199

Ogg, R. J., Zou, P., Allen, D. N., Hutchins, S. B., and Dutkiewicz, R. M., Mulhern, R. K. (2008). Neural correlates of a clinical continuous performance test. *Magn. Reson. Imaging* 26, 504–512. doi: 10.1016/j.mri.2007.09.004

Oosterman, J. M., Vogels, R. L., van Harten, B., Gouw, A. A., Poggesi, A., Scheltens, P., et al. (2010). Assessing mental flexibility: neuroanatomical and neuropsychological correlates of the Trail Making Test in elderly people. *Clin. Neuropsychol.* 24, 203–219. doi: 10.1080/13854040903482848

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., and Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia* 28, 1021–1034. doi: 10.1016/0028-3932(90)90137-D

Pantelis, C., and Brewer, W. (1996). "Neurocognitive and neurobehavioural patterns and the syndromes of schizophrenia: role of frontal-subcortical networks," in Schizophrenia: *A Neuropsychological Perspective*, eds C. Pantelis, H. E. Nelson, and T. R. E. Barnes (West Sussex: John Wiley and Sons Ltd), 317-343.

Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342:1238411. doi: 10.1126/science.1238411

Park, S., and Gooding, D. C. (2014). Working memory impairment as an endophenotypic marker of a schizophrenia diathesis. *Schizophr. Res. Cogn.* 1, 127–136. doi: 10.1016/j.scog.2014.09.005

Pérez-Iglesias, R., Tordesillas-Gutiérrez, D., McGuire, P. K., Barker, G. J., Roiz-Santiañez, R., Mata, I., et al. (2010). White matter integrity and cognitive impairment in first-episode psychosis. *Am. J. Psychiatry* 167, 451–458. doi: 10.1176/appi.ajp.2009.09050716

Płotek, W., Łyskawa, W., Kluzik, A., Grześkowiak, M., Podlewski, R., Zaba Z, and Drobnik, L. (2014). Evaluation of the Trail Making Test and interval timing as measures of cognition in healthy adults: comparisons by age, education, and gender. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res* 20, 173–181. doi: 10.12659/MSM.889776

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* 72, 692–697. doi: 10.1016/j.neuron.2011.11.001

Pöppel, E. (1989). "Taxonomy of the subjective: an evolutionary perspective," in *Neuropsychology of Visual Perception*, ed J. W. Brown (Hillsdale, NJ: Lawrence Erlbaum Associates), 219–232.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., et al. (2011). Functional network organization of the human brain. *Neuron* 72, 665–678. doi: 10.1016/j.neuron.2011.09.006

Preiss, M., Bartoš, A., Cermáková, R., Nondek, M., Benešová, M., Rodriguez, M., et al. (2012). *Neuropsychologická baterie Psychiatrického centra Praha. Klinické vyšetrení základních kognitivních funkcí. 3. Vydání.* Praha: Psychiatrické centrum Praha.

Ragland, J. D., Gur, R. C., Valdez, J., Turetsky, B. I., Elliott, M., Kohler, C., et al. (2004). Event-related fMRI of frontotemporal activity during word encoding and recognition in schizophrenia. *Am. J. Psychiatry* 161, 1004–1015. doi: 10.1176/appi.ajp.161.6.1004

Ragland, J. D., Moelter, S. T., McGrath, C., Hill, S. K., Gur, R. E., Bilker, W. B., et al. (2003). Levels-of-processing effect on word recognition in schizophrenia. *Biol. Psychiatry* 54, 1154–1161. doi: 10.1016/S0006-3223(03)00235-X

Ramage, A., Bayles, K., Helm-Estabrooks, N., and Cruz, R. (1999). Frequency of perseveration in normal subjects. *Brain Lang.* 66, 329–340. doi: 10.1006/brln.1999.2032

Riccio, C. A., Reynolds, C. R., Lowe, P., and Moore, J. J. (2002). The continuous performance test: a window on the neural substrates for attention?. *Arch. Clin. Neuropsychol.* 17, 235–272. doi: 10.1093/arclin/17.3.235

Rodriguez, M., Spaniel, F., Konradova, L., Sedlakova, K., Dvorska, K., Prajsova, J., et al. (2015). Comparison of visuospatial and verbal abilities in first psychotic episode of schizophrenia spectrum disorder: impact on global functioning and quality of life. *Front. Behav. Neurosci.* 9:322. doi: 10.3389/fnbeh.2015.00322

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome Jr, E. D., and Beck, L. H. (1956). A continuous performance test of brain damage. *J. Consult. Psychol.* 20:343. doi: 10.1037/h0043220

Rouleau, I., Imbault, H., Laframboise, M., and Bédard, M. A. (2001). Pattern of intrusions in verbal recall: comparison of Alzheimer's disease, Parkinson's disease, and frontal lobe dementia. *Brain Cogn.* 46, 244–249. doi: 10.1016/S0278-2626(01)80076-2

Rueckert, L., and Grafman, J. (1996). Sustained attention deficits in pat ients with right frontal lesions. *Neuropsychologia* 34, 953–963. doi: 10.1016/0028-3932(96)00016-4

Ruffolo, L. F., Guilmette, T. J., and Willis, G. W. (2000). FORUM comparison of time and error rates on the trail making test among patients with head injuries, experimental malingerers, patients with suspect effort on testing,

and normal controls. *Clin. Neuropsychol.* 14, 223–230. doi: 10.1076/1385-4046(200005)14:2;1-Z;FT223

Salgado-Pineda, P., Junqué, C., Vendrell, P., Baeza, I., Bargalló, N., Falcón, C., et al. (2004). Decreased cerebral activation during CPT performance: structural and functional deficits in schizophrenic patients. *Neuroimage* 21, 840–847. doi: 10.1016/j.neuroimage.2003.10.027

Sánchez-Cubillo, I., Perianez, J. A., Adrover-Roig, D., Rodriguez-Sanchez, J. M., Rios-Lago, M., Tirapu, J., et al. (2009). Construct validity of the Trail Making Test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc.* 15, 438–450. doi: 10.1017/S1355617709090626

Sanfilipo, M., Lafargue, T., Rusinek, H., Arena, L., Loneragan, C., Lautin, A., et al. (2002). Cognitive performance in schizophrenia: relationship to regional brain volumes and psychiatric symptoms. *Psychiatry Res. Neuroimaging* 116, 1–23. doi: 10.1016/S0925-4927(02)00046-X

Savage, C. R., Deckersbach, T., Heckers, S., Wagner, A. D., Schacter, D. L., Alpert, N. M., et al. (2001). Prefrontal regions supporting spontaneous and directed application of verbal learning strategies: evidence from PET. *Brain* 124, 219–231. doi: 10.1093/brain/124.1.219

Schindler, A. G., Caplan, L. R., and Hier, D. B. (1984). Intrusions and perseverations. *Brain Lang.* 23, 148–158. doi: 10.1016/0093-934X(84)90013-0

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A Handbook.* Los Angeles, CA: Western Psychological Services.

Schnider, A., von Däniken, C., and Gutbrod, K. (1996). The mechanisms of spontaneous and provoked confabulations. *Brain* 119, 1365–1375. doi: 10.1093/brain/119.4.1365

Schouten, E. A., Schiemanck, S. K., Brand, N., and Post, M. W. (2009). Long-term deficits in episodic memory after ischemic stroke: evaluation and prediction of verbal and visual memory performance based on lesion characteristics. *J. Stroke Cerebrovasc. Dis.* 18, 128–138. doi: 10.1016/j.jstrokecerebrovasdis.2008.09.017

Schuepbach, D., Egger, S., and Herpertz, S. C. (2016). Cerebral hemodynamics in schizophrenia during the Trail Making Test: a functional transcranial Doppler sonography study. *Eur. Psychiatry* 33:S107. doi: 10.1016/j.eurpsy.2016.01.094

Seltzer, J., Conrad, C., and Cassens, G. (1997). Neuropsychological profiles in schizophrenia: paranoid versus undifferentiated distinctions. *Schizophr. Res.* 23, 131–138. doi: 10.1016/S0920-9964(96)00094-1

Senden, M., Reuter, N., van den Heuvel, M., Goebel, R., Gustavo, D., and Gilson, M. (2018). Task-related effective connectivity reveals that the cortical rich club gates cortex-wide communications. *Hum. Brain Mapp.* 39, 1246–1262. doi: 10.1002/hbm.23913

Sepede, G., Spano, M. C., Lorusso, M., De Berardis, D., Salerno, R. M., Di Giannantonio, M., et al. (2014). Sustained attention in psychosis: neuroimaging findings. *World J. Radiol.* 6:261. doi: 10.4329/wjr.v6.i6.261

Shao, Z., Janse, E., Visser, K., and Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* 5:772. doi: 10.3389/fpsyg.2014.00772

Shaposhnyk, V., and Villa, A. E. P. (2012). "The effect of short and long reciprocal projections on evolution of hierarchical neural networks at ICANN 2012," in *Artificial Neural Networks and Machine Learning ICANN 2012* (Berlin; Heidelberg), 371–378.

Sheffield, J. M., and Barch, D. M. (2016). Cognition and resting state functional connectivity in schizophrenia. *Neurosci. Biobehav. Rev.* 61, 108–120. doi: 10.1016/j.neubiorev.2015.12.007

Silverstein, S. M. (2008). Measuring specific, rather than generalized, cognitive deficits and maximizing between-group effect size in studies of cognition and cognitive change. *Schizophr. Bull.* 34, 645–655. doi: 10.1093/schbul/sbn032

Singer, W. (2013). Cortical dynamics revisited. *Trends Cogn. Sci.* 17, 616–626. doi: 10.1016/j.tics.2013.09.006

Smith, S. M., Miller, K. L., Moeller, S., Xu, J., Auerbach, E. J., Woolrich, M. W., et al. (2012). Temporally-independent functional modes of spontaneous brain activity. *Proc. Natl. Acad. Sci.* 109, 3131–3136. doi: 10.1073/pnas.1121329109

Sporns, O. (2013). The human connectome: origins and challenges. *Neuroimage* 80, 53–61. doi: 10.1016/j.neuroimage.2013.03.023

Sridharan, D., Levitin, D. J., and Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc. Natl. Acad. Sci.* 105, 12569–12574. doi: 10.1073/pnas.0800005105

Stephan, K. E., and Friston, K. J. (2010). Analyzing effective connectivity with fMRI. *Wiley Interdisc. Rev. Cogn. Sci.* 1, 446–459. doi: 10.1002/wcs.58

Stephan, K. E., Mattout, J., David, O., and Friston, K. J. (2006). Models of functional neuroimaging data. *Curr. Med. Imaging Rev.* 2, 15–34. doi: 10.2174/157340506775541659

Stevens, M. C., Kiehl, K. A., Pearlson, G. D., and Calhoun, V. D. (2009). Brain Networks dynamics during error commission. *Hum. Brain Mapp.* 30, 24–37. doi: 10.1002/hbm.20478

Stirling, J. D., Hellewell, J. S., and Quraishi, N. (1998). Self-monitoring dysfunction and the schizophrenic symptoms of alien control. *Psychol. Med.* 28, 675–683. doi: 10.1017/S0033291798006679

Strauss, E., Sherman, E. M., and Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary.* New York, NY: Oxford University Press.

Stuss, D. T., Bisschop, S. M., Alexander, M. P., Levine, B., Katz, D., and Izukawa, D. (2001). The trail making test: a study in focal lesion patients. *Psychol. Assess.* 13:230. doi: 10.1037/1040-3590.13.2.230

Subotnik, K. L., Nuechterlein, K. H., Green, M. F., Horan, W. P., Nienow, T. M., Ventura, J., et al. (2006). Neurocognitive and social cognitive correlates of formal thought disorder in schizophrenia patients. *Schizophr. Res.* 85, 84–95. doi: 10.1016/j.schres.2006.03.007

Taki, Y., Kinomura, S., Sato, K., Goto, R., Wu, K., Kawashima, R., et al. (2011). Correlation between gray/white matter volume and cognition in healthy elderly people. *Brain Cogn.* 75, 170–176. doi: 10.1016/j.bandc.2010.11.008

Takizawa, R., Kasai, K., Kawakubo, Y., Marumo, K., Kawasaki, S., Yamasue, H., et al. (2008). Reduced frontopolar activation during verbal fluency task in schizophrenia: a multi-channel near-infrared spectroscopy study. *Schizophr. Res.* 99, 250–262. doi: 10.1016/j.schres.2007.10.025

Taylor, R., Théberge, J., Williamson, P. C., Densmore, M., and Neufeld, R. W. (2016). ACC neuro-over connectivity is associated with mathematically modeled additional encoding operations of schizophrenia Stroop-task performance. *Front. Psychol.* 7:1295. doi: 10.3389/fpsyg.2016.01295

Thompson, J. C., Stopford, C. L., Snowden, J. S., and Neary, D. (2005). Qualitative neuropsychological performance characteristics in frontotemporal dementia and Alzheimer's disease. *J. Neurol. Neurosurg. Psychiatry* 76, 920–927. doi: 10.1136/jnnp.2003.033779

Titova, O. E., Lindberg, E., Elmståhl S., Lind, L., Schiöth, H. B., and Benedict, C. (2016). Association between shift work history and performance on the trail making test in middle-aged and elderly humans: the EpiHealth study. *Neurobiol. Aging* 45, 23–29. doi: 10.1016/j.neurobiolaging.2016.05.007

Townsend, J. T., and Ashby, F. G. (1983). *The Stochastic Modeling of Elementary Psychological Processes.* New York, NY: Cambridge University Press.

Turetsky, B. I., Moberg, P. J., Mozley, L. H., Moelter, S. T., Agrin, R. N., Gur, R. C., et al. (2002). Memory-delineated subtypes of schizophrenia: relationship to clinical, neuroanatomical, and neurophysiological measures. *Neuropsychology* 16:481. doi: 10.1037/0894-4105.16.4.481

van Buuren, M., Gladwin, T. E., Zandbelt, B. B., Kahn, R. S., and Vink, M. (2010). Reduced functional coupling in the default-mode network during self-referential processing. *Hum. Brain Mapp.* 31, 1117–1127. doi: 10.1002/hbm.20920

van den Heuvel, M. P., and Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn. Sci.* 17, 683–696. doi: 10.1016/j.tics.2013.09.012

Van Dijk, K. R., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., and Buckner, R. L. (2010). Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *J. Neurophysiol.* 103, 297–321. doi: 10.1152/jn.00783.2009

Van Vreeswijk, C., Abbott, L. F., and Ermentrout, G. B. (1994). When inhibition not excitation synchronizes neural firing. *J. Comput. Neurosci.* 1, 313–321. doi: 10.1007/BF00961879

Vandevelde, A., Leroux, E., Delcroix, N., and Dollfus, S. (2017). Fronto-subcortical functional connectivity in patients with schizophrenia and bipolar disorder during a verbal fluency task. *World J. Biol. Psychiatry* 1–9. doi: 10.1080/15622975.2017.1349339

Volz, H. P., Gaser, C., Häger, F., Rzanny, R., Pönisch, J., Mentzel, H.-J., et al. (1999). Decreased frontal activation in schizophrenics during stimulation with the continuous performance test-a functional magnetic resonance imaging study. *Eur. Psychiatry* 14, 17–24. doi: 10.1016/S0924-9338(99)80711-1

Wagner, S., Sebastian, A., Lieb, K., Tüscher, O., and Tadić, A. (2014). A coordinate-based ALE functional MRI meta-analysis of brain activation during verbal fluency tasks in healthy control subjects. *BMC Neurosci.* 15:19. doi: 10.1186/1471-2202-15-19

Wang, X., Zhang, W., Sun, Y., and Chen, A. (2016). Aberrant intra-salience network dynamic functional connectivity impairs large-scale network interactions in schizophrenia. *Neuropsychologia* 93, 262–270. doi: 10.1016/j.neuropsychologia.2016.11.003

Weiss, E. M., Hofer, A., Golaszewski, S., Siedentopf, C., Brinkhoff, C., Kremser, C., et al. (2004). Brain activation patterns during a verbal fluency test—a functional MRI study in healthy volunteers and patients with schizophrenia. *Schizophr. Res.* 70, 287–291. doi: 10.1016/j.schres.2004.01.010

Weiss, E. M., Ragland, J. D., Brensinger, C. M., Bilker, W. B., Deisenhammer, E. A., and Delazer, M. (2006). Sex differences in clustering and switching in verbal fluency tasks. *J. Int. Neuropsychol. Soc.* 12, 502–509. doi: 10.1017/S1355617706060656

Weiss, E. M., Siedentopf, C., Hofer, A., Deisenhammer, E. A., Hoptman, M. J., Kremser, C., et al. (2003). Brain activation pattern during a verbal fluency test in healthy male and female volunteers: a functional magnetic resonance imaging study. *Neurosci. Lett.* 352, 191–194. doi: 10.1016/j.neulet.2003.08.071

Whitaker, H. (1972).*Phonological Processes and Brain Mechanisms*. Springer-Verlag.

White, C. N., Mumford, J. A., and Poldrack, R. A. (2012). Perceptual criteria in the human brain. *Th. J. Neurosci.* 32, 16716–16724. doi: 10.1523/JNEUROSCI.1744-12.2012

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., et al. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proc. Natl. Acad. Sci.* 106, 1279–1284. doi: 10.1073/pnas.0809141106

Wierenga, L. M., van den Heuvel, M., van Dijk, S., Rijks, Y., de Reus, M. A., and Durston, S. (2015). The development of brain network architecture. *Hum. Brain Mapp.* 37, 717–729. doi: 10.1002/hbm.23062

Wolfe, N., Linn, R., Babikian, V. L., Knoefel, J. E., and Albert, M. L. (1990). Frontal systems impairment following multiple lacunar infarcts. *Arch. Neurol.* 47, 129–132. doi: 10.1001/archneur.1990.00530020025010

Wolk, D. A., and Dickerson, B. C. (2011). Fractionating verbal episodic memory in Alzheimer's disease. *Neuroimage* 54, 1530–1539. doi: 10.1016/j.neuroimage.2010.09.005

Woods, D. L., Herron, T. J., Yund, E. W., Hink, R. F., Kishiyama, M. M., and Reed, B. (2011). Computerized analysis of error patterns in digit span recall. *J. Clin. Exp. Neuropsychol.* 33, 721–734. doi: 10.1080/13803395.2010.550602

Yang, Z., Jutagir, D. R., Koyama, M. S., Craddock, R. C., Yan, C.-G., Shehzad, Z., et al. (2015). Intrinsic brain indices of verbal working memory capacity in children and adolescents. *Dev. Cogn. Neurosci.* 15, 67–82. doi: 10.1016/j.dcn.2015.07.007

Yechiam, E., Goodnight, J., Bates, J. E., Busemeyer, J. R., Dodge, K. A., Pettit, G. S., et al. (2006). A formal cognitive model of the Go/No-Go Discrimination Task: evaluation and implications. *Psychol. Assess.* 18, 239–249. doi: 10.1037/1040-3590.18.3.239

Ystad, M., Eichele, T., Lundervold, A. J., and Lundervold, A. (2010). Subcortical functional connectivity and verbal episodic memory in healthy elderly—a resting state fMRI study. *Neuroimage* 52, 379–388. doi: 10.1016/j.neuroimage.2010.03.062

Zakzanis, K. K., Mraz, R., and Graham, S. J. (2005). An fMRI study of the trail making test. *Neuropsychologia* 43, 1878–1886. doi: 10.1016/j.neuropsychologia.2005.03.013

Zaytseva, Y., Chan, R. C., Pöppel, E., and Heinz, A. (2015). Luria revisited: cognitive research in schizophrenia, past implications and future challenges. *Philos. Ethics Humanit. Med.* 10:4. doi: 10.1186/s13010-015-0026-9

Embodied Decision-Making Style: Below and Beyond Cognition

Brenda L. Connors[1]* and Richard Rende[2]

[1] Office of the President, Naval War College, Newport, RI, United States, [2] Social Behavioral Research Applications, Phoenix, AZ, United States

There is growing recognition of the essential role of sensorimotor processes as not just a supporter of the cognitive aspects of decision making, but rather as a foundation for all the coordinated physical and mental activities that go into how we make decisions. We illuminate concepts and methods for examining embodied decision making through the lens of Movement Pattern Analysis (MPA). MPA is as a prime example of a conceptually rooted observational methodology for deciphering embodied decision making and for decoding how people differ as decision makers with respect to cognitive motivational priorities. The historical origins of MPA that predated the formalized recognition of embodied cognition are presented, along with an overview of both the theoretical model and methodology. Advances in research on two psychometric benchmarks of observational research—inter-rater reliability and predictive validity—are highlighted as an empirical platform for the strong promise of MPA as a tool for understanding individual differences in embodied decision-making style. Future directions for research are considered—specifically with respect to the potential for utilizing automated coding, and the need for collaborative neuroscience research efforts—which would support further understanding of how decoding movement patterning captures human motivation at the level of sensory, motoric, cognitive and action integration which drives how people function as decision makers.

Keywords: embodied cognition, decision-making style, decision-making process, organizational decision making, movement pattern analysis, human movement science, leadership analysis, leadership development

DECISION MAKING AS EMBODIED COGNITION

In non-academic circles, when we talk about decision making, our language gravitates to discussions of thought processes. We reference how we weigh options, if we waffle or pull the trigger, if decisions are no-brainers or extremely hard. The emphasis is on the "decision" as something we do with our mind, the range of decisions we are faced with, and the different types of thinking that help us make decisions.

Much of the classic literature on decision making has taken a similar perspective. In terms of theory, decision making has been dissected into a number of cognitive processes. Research has brought much insight into how these processes are utilized to make decisions, typically under different types of demands, ranges of complexity, and varying rewards and consequences (e.g., Connors et al., 2013, 2018a).

While we have learned much about the cognitive side of decision making, there are still major gaps to be filled in terms of understanding how people enact decisions in real world settings. One key consideration is that decision making, like other forms of cognition, actually involves much more than cogitation. Here research on decision making is aligned

with the thrust of work linking the body to thought, as articulated across social science domains and collectively known as embodied cognition. That said, while there are nearly two decades of scholarly papers devoted, formally, to embodied cognition, most do not focus explicitly on decision making.

That is starting to change, and for good reason. People are biologically wired not just to think, but also to determine when, why, and how to act in relation to changing internal and external physical, emotional and social surroundings. There is an inherent "sensorimotor coupling" between person and environment that defines the essence of embodied cognition[1], as expressed as changes in observable indicators such as facial expressions, postures, and gestures (Pietrazak et al., 2018). With respect to decision making, we are seeing a current of recognition of the essential role of sensorimotor processes as not just a supporter of cognition, but rather as a foundation for all the coordinated physical and mental activities that go into how we make decisions (Connors et al., 2018a). For example:

> "The central statement of embodied choice is the existence of bidirectional influences between action and decisions. This implies that … the action dynamics and its constraints (e.g., current trajectory and kinematics) influence the decision making process" (Lepora and Pezzulo, 2015, p. 1).

In this paper, we illuminate concepts and methods for examining embodied decision making through the lens of decoding signature movement patterns. We focus on Movement Pattern Analysis (MPA)—including its historical origins that predated the formalized recognition of embodied cognition—which serves as a prime example of a conceptually rooted methodology for deciphering embodied decision making.

DECODING MOVEMENT AS INSIGHT INTO EMBODIED DECISION MAKING

Our point of departure is to focus intensively on direct observation of the body's patterning of movement to provide unique insight into decision making. We suggest that human movement underlies the inherent connection between thinking and behavior and resides "below and beyond" cognition. Within the sphere of embodied cognition, it has been well appreciated that movement supports cognitive functions (and hence is "below" cognition) by being a mechanism responsible for orienting the body to take in sensory information and facilitate perception, and to translate thought into action. That said, we wish to promote a deeper principle—that patterning of movement captures *human motivation* at the level of sensory, motoric, cognitive and action organization that drives how people go about making decisions (and hence is "beyond" cognition).

This principle has long roots and dates to the innovative insights of pioneers in movement analysis as it relates to

human cognition and behavior, which we have reviewed in depth (Connors et al., 2018a) and hence summarize here. The Hungarian polymath Rudolf Laban (1878-1958) was a movement theorist and father of movement analysis and notation. A visual artist as well as dancer and choreographer, his acute observations of dance led him to decipher how movement conveys inner attitudes and the expression of psychophysical and emotional cues (Laban, 1950). Importantly, this led him to devise systematic ways to decode movement via two systems of analysis (which we would now call observational coding systems). Labanotation or *Kinetography Laban* is a notation system for recording and analyzing human movement. Analogous to music notation, it records complex actions of the body and dynamic nuance via symbols. Notating the movement and expression provided a way for dance choreography to be reproduced from a written score (Laban, 1928).

Notably, Laban also created Laban Movement Analysis (LMA), which is a multidisciplinary method and language for describing, visualizing, interpreting and documenting all varieties of human movement. LMA is particularly relevant for the purposes of this paper, as it eventually became an innovative platform for observing decision making via the coding of movement in naturalistic settings (Moore, 2005; Lamb, 2012; Connors et al., 2018a). In brief, in 1941 Laban was in England (as a refugee from Nazi Germany), where he was recruited to collaborate with F. C. Lawrence (an engineer and time motion expert) to increase productivity of women on factory assembly lines (who were working there as men were serving in the war). Laban pioneered an "Industrial Rhythm" approach that apprehended and honored the unique rhythmic patterns of workers, which yielded many positive results in the factory. Importantly, Laban and Lawrence (1947) also expanded the method to study clerical and managerial workers as they performed their duties. Here they detected distinctive movement patterns that corresponded to different types of white collar jobs and tasks that would not typically be considered to be physical in nature, providing, in the naturalistic setting, an insight and perspective consistent with embodied decision making.

It is particularly important to recognize that the detection of movement patterns came from observations of individuals moving freely in their naturalistic (work) environment. This provided an authenticity—what we would refer to today as ecological validity—that drove him to devise meaningful systems for recording movement and notating a number of complex patterns that corresponded to psychological processes. This point also reinforces the role of the highly trained and sophisticated human observer, who is capable of parsing what is apprehended as observable into systematic recording systems that capture the deeper significance of movement patterning. We will return to these two points later in this paper.

MOVEMENT PATTERN ANALYSIS (MPA): EMBODIED DECISION MAKING

One of Laban's protégés, Warren Lamb, was brought in and was trained to examine in more depth the relations between movement patterns and job responsibilities. Lamb

---

[1]"Cognition is embodied when it is deeply dependent upon features of the physical body of an agent, that is, when aspects of the agent's body beyond the brain play a significant causal or physically constitutive role in cognitive processing": *The Stanford Encyclopedia of Philosophy*, plato.stanford.edu/entries/embodied-cognition.

continued this work and observed how distinctive patterns of movement corresponded to higher-order functions, including decision-making processes. This led Lamb to formalize his intensive observations by developing a conceptual model and corresponding coding system for recording signature movements that align with stages of decision making, collectively referred to as MPA. Lamb's approach included a number of notable features (e.g., Moore, 2005; Lamb, 2012; Connors et al., 2018a), each of which illuminate his prescient abilities as an observer of human movement as they are key elements of current thinking in decision-making science.

First, Lamb's observations led him to conceptualize decision making as involving three *distinct stages*. These are as follows:

- *Attending*: Engaging in the preliminary consideration (e.g., of information deeply and broadly) that goes into making a decision knowledgeably.
- *Intending*: Coming to a resolute ascertainment of the best course of action driving the decision to be made.
- *Committing*: Mobilizing to implement both tactically and in longer term staging what has been intended; initiating and pacing action adroitly to attain goals and avoid pitfalls.

Lamb also proposed that, within each stage, individuals have a need to coordinate two complementary goals, or motivations, that drive the overall decision-making process. These are known as the Overall Factors in the MPA model, and are as follows:

- *Assertion*: Applying effort or elbow grease by focusing, pushing and pacing tactically to make the decision happen.
- *Perspective*: Positioning the body to create an environment in which to reflect on the broader strategic context and how parts fit together.

The approach observes and records signature movement patterns that corresponded to each stage in the MPA model, and which reflect either Assertion or Perspective. Lamb focused on a specific movement phenomenon that was the platform for the body to engage in a range of decision-making processes—the Posture-Gesture Merger, or *PGM* (Lamb, 2012). A PGM reflects the coherent integration of both a gesture (an action that is isolated to one or two body parts, such as head nodding or a foot and thigh crossing) along with a postural action involving all body parts (such as the whole body moving in a head-to-toe jump to get someone's attention) (Moore, 2005; Lamb, 2012; Connors et al., 2018a).

The MPA model identifies PGMs that are organized along the two interrelated Overall Factors in relation to the three planes of motion (horizontal, vertical, and sagittal) (Connors et al., 2018a), They are hence decoded to reveal how a person effortfully applies energy to investigate, push or pace (Assertion) vs. how in the kinesphere around the body a person shapes to explore, prioritize, and anticipate (Perspective). PGMs have been shown to be generated spontaneously by individuals in conjunction with verbalizations that reflect authenticity (Winter et al., 1989; Winter, 1992; Lamb, 2012), suggesting that they are reflective of meaningful thought and action.

Lamb discovered that there are a variety of PGMs that correspond to the stages in the MPA decision-making model, and

which align with either Assertion or Perspective. He continually refined this discovery diving deeper into its essence of the two Overall Factors and the three stages of decision making and uncovered six distinct Action Motivations which reflect the two types of Overall Factors at each stage, one motivation reflecting Assertion, and one motivation reflecting Perspective. Each of the Action Motivations can be discerned by the highly trained observer by detecting signature PGMs (see Connors et al., 2013, for more details):

- During Attending, individuals engage in *Investigating* (applying effort, or hence Assertion, to scan/probe/analyze relevant information within a prescribed area) and *Exploring* (positioning oneself, to gain hence Perspective, and to be open to a wide range of ideas along with potentially relevant resources). In terms of PGMs, an example of Investigating would be an individual zeroing in on a target (e.g., a map) with head and torso, merged with a pointing gesture; an example of Exploring would be spreading the chest and upper torso broadly while having that movement flow into an arm gesture.
- During Intending, individuals can engage in *Determining* (devoting effort, or hence Assertion, to build the resolve necessary to formulate a position and take a stand) and *Evaluating* (positioning oneself, to gain Perspective by sizing up options and setting priorities). In terms of PGMs, an example of Determining would be pressing the body (including hands and feet) down in a chair followed by and merging into gesturing with the chin; an example of Evaluating would be to rise up on one side of the body, crossing the legs and flowing into the "thinker" pose with arms crossed and one hand supporting the head.
- During Committing, individuals can engage in *Timing* (applying effort, or hence Assertion, to implement the decision at the right moment) and *Anticipating* (positioning oneself to gain Perspective in a strategic way to guide and monitor the implementation of the decision staging). In terms of PGMs, an example of Timing would be a very quick forearm and hand gesture followed by the whole body shifting rapidly; and example of Anticipating would be advancing the chest forward and having that movement flow into an head gesture in a meandering way.

The MPA model thus provides a detailed framework for understanding embodied decision making by articulating the distinct stages and observable movement patterns that reflect how an individual interacts with the motion factor and engages in complementary management decision-making processes during each stage.

MOVEMENT PATTERN ANALYSIS (MPA): EMBODIED DECISION MAKERS

Importantly, the MPA model devised by Lamb also offered futuristic insight into how observation of movement patterning could help us understand *decision makers*, not just decision making. Lamb discerned that while the MPA model identifies universal mechanisms by which we engage in embodied decision

making, individuals differ in the extent to which they prioritize and sequence the specific processes. In terms of the stages, people vary in how much they "favor" Attending, Intending and Committing (e.g., some may be "high committers," whereas others may be "low committers;" some may begin with Committing while others may nearly skip over implementing all together). A fundamental window into how people differ as decision makers comes from how they achieve a balance between the complementary Overall Factors of Assertion and Perspective. Lamb postulated that how each individual achieves their own personal mix of Assertion and Perspective—within and across stages—reveals their cognitive motivational predispositions as a decision maker.

As such, MPA serves as a unique method for capturing *decision-making style* as it is used to provide insight into how each individual goes about making decisions. This is style in the sense that it is not focused on what decisions people make, but rather the way they go about navigating them in terms of balancing Assertion and Perspective across the distinct stages of decision making. MPA has been used in this way for over 50 years by organizations to understand decision makers as leaders, guide selection and placement of top personnel, and inform the building of management teams (Connors et al., 2018a). It has provided the type of ecological validity that Laban and Lawrence (1947) achieved in their pioneering work in the factory and the office.

For the purposes of this paper, we focus on more recent efforts to determine, through the lens of research, the utility of MPA as a method for decoding decision-making style. The vast majority of methods that have been used, in the field, to measure decision-making style are rooted in self-assessment, in the form of questionnaires and inventories (Connors et al., 2013, 2016). While self-perception is an important aspect of decision-making style, there are limits to it; for example, there may be disconnects between how people perceive themselves as decision makers, and how they function as decision makers. Furthermore, there is much interest in moving beyond self-report by focusing on observational approaches that can dig deeper into decision-making processes and pinpoint with more precision how people differ as decision makers (Connors et al., 2016).

Such a consideration is especially important in terms of the types of scenarios that decision makers face. Consider, for example, leaders who navigate complex and ambiguous decision-making situations including crises, and orchestrate time-sensitive and high-stakes negotiations. A key idea here is that there is no one optimal decision-making style and no one way to make a decision; leaders who are high in Assertion or high in Perspective can be equally effective in the same position. That said, having methods that can help leaders understand themselves as decision makers would undoubtedly help them optimize their performance along with improving how they function with their colleagues (e.g., by including decision makers on their team who offer a complementary decision-making style in order to provide better overall balance on issues and crises).

While the long history of MPA in application to organizations certainly serves as proof of principle of its utility, even more confidence in the method would be gained via empirical study.

To that end, we now turn to summarizing recent advances in evaluating core psychometric properties of MPA, particularly benchmark indicators including reliability and validity that are evaluated for any system of measurement rooted in observation (e.g., Dishion and Synder, 2004; Rende et al., 2009; Slomkowski et al., 2009; Bakeman and Quera, 2011; Girard and Cohn, 2016).

IS MPA RELIABLE ACROSS EXPERT OBSERVERS?

Highlighting the many intellectual and methodological contributions of visionaries such as Laban and Lamb illuminates the complexity of mastering MPA coding.

Applying MPA requires decoding integrated patterns of motion infused with degrees of pressure, acceleration, focus, and energy that are manifested in physical space and distributed across temporal parameters in a continuous and spontaneous display throughout the whole body. Doing this right is a hard-earned skill that goes beyond detecting isolated and truncated behaviors. Recognizing overall patterns of movement demands acuity at coupling quantitative and qualitative measures that are sensitive enough to offer discrimination across a wide range of decision-making styles.

To this end, MPA analysts go through rigorous training to develop the expertise necessary to skillfully decode signature movement patterns, and are certified as such. Training includes mastering an advanced level of observational expertise as well as learning how to administer the standard protocol to conduct a full MPA profile, which is executed in a semi-structured interview. The goal is to engage the interviewee in a discussion of his or her career and biographical history, which provides a naturalistic platform to observe their embodied decision-making motivations and priorities. The interview typically requires 90–120 min to allow the analyst to gather a sufficient sampling of PGMs to provide confidence in establishing an individual's signature patterning with respect to the stages of decision making, Overall Factors, and Action Motivations. Coding is done during the observation in real time, and can also be undertaken via review of a videotaped recording of the interview.[2] Importantly, the certified MPA analyst is also trained to determine, much like a clinician, if and when they have observed a sufficient sampling of PGMs to feel confident that they have acquired a representative baseline for determining an individual's propensities. A final aspect of the assessment process is interpreting and writing in words the findings with a refined degree of balancing the many variations in movement that are quantified in the coding into a coherent picture of propensity and pattern.

The total number of PGMs provides a "denominator" for each individual, such that their movement patterning can be represented as percentages out of a grand total of 100 per cent reflecting the PGMs expressed as Assertion or Perspective within and across the stages of the MPA model (as reflected in the Action

---

[2]MPA can also be applied to videotaped recordings of an individual as captured in naturalistic settings, if the analyst can decipher a sufficient sampling of PGMs to support the crafting of the signature patterning of that individual.

Motivations). This self-referenced balancing in percentage terms provides an elegant quantitative representation of their decision-making priorities and motivations.

Methods for Gauging Inter-rater Reliability of MPA

Given the richness and specificity of MPA, a key issue, and a requirement for any observational system, is gauging the inter-rater reliability of MPA, which informs us of the extent to which different trained experts converge on their assessments of an individual's decision-making profile. In general, gauging reliability can be a formidable task when the methodology demands high precision in separating signal from noise in the stream of real-time behavior, as is the case with MPA. Moreover, MPA is unlike many other nonverbal behavioral coding systems that specialize primarily within one subsystem and code all relevant movement behaviors in that area (e.g., gesture or facial expression). Indeed, MPA is distinctive in that two areas of movement repertoire or subsystems—namely the posture and gesture of movement—are involved that ultimately result in a merged quality throughout the entire body. These synchronized movements merge and flood into a continuous stream of energetic activity from head-to-toe.

The implication is that calculating inter-rater reliability of MPA is not just a straightforward exercise of determining if different raters agree when coding the behaviors of interest, namely, the categories of PGMs. One standard approach used in observational research would be to measure the degree to which different coders agreed on every rating that was made during the observation with a degree of time anchoring (e.g., some window of time applied to any event detected by a rater, during which time another rater would be in agreement if they also recorded the event). Here percent agreement across raters would be calculated, and inter-rater reliability would be calculated using indicators that correct for chance agreement, such as Cohen's Kappa. An alternative would be to aggregate totals of PGMs within categories and then assess inter-rater reliability on these quantitative measures using the Intraclass Correlation Coefficient (ICC). An issue, however, is that neither of these conventional strategies would be aligned with the nub of MPA coding. MPA is oriented toward the patterning of target behaviors *within each subject* (rather than total counts of behaviors), which is thought to best reveal decision-making propensities.

We have framed such considerations as an empirical issue, as we compare the level of reliability of coding individual patterning of signature movements, to the level of reliability for raw count tabulations. For each individual, we would have a *total* number of PGMs coded as reflecting either Assertion or Perspective (raw counts), as well as a *proportional representation* for each subject that captures the balance between these Overall Factors relative to each individual's total number of PGMs (individual patterning). A hypothetical serves to make the issue more transparent. Consider two individuals: subject one who had 100 PGMs coded as Assertion, and subject two who had 120 PGMs coded as Assertion. Based on these numbers, it is not possible to conclude that subject two is more inclined to Assertion as

compared to subject one. What is missing is the denominator, or the total count of all PGMs (those coded as either Assertion or Perspective). For example, if subject one had a tally total of 140 PGMs, then about 70% of this individual's PGMs (100/120) would indicate Assertion. If subject two had a total tally of 210 PGMs coded, then nearly 60% of this individual's PGMs would indicate Assertion (120/210), which is lower than the 70% for subject one. We would thus conclude, in terms of individual patterning, that subject one was more inclined to Assertion than subject two, even though subject two had a higher absolute number of Assertion PGMs.

Using this approach, we computed inter-rater reliability using two different indicators of how each individual expressed Assertion and Perspective in their movement patterning. One indicator focused on the *relative percent* of Perspective and Assertion as referenced by each individual subject's own baseline and as such labeled as a P/A Balance Score (Connors et al., 2014). The P/A Balance Score is calculated as [% Perspective – % Assertion] and conveys the conceptual point that MPA determines the relative importance that a person allocates to each of the Overall Factors, which is displayed in how they move unconsciously to each motivation factor as a decision maker. A value greater than zero indicates more emphasis on Perspective; a value less than zero represents more emphasis on Assertion; and zero signals equal emphasis on both Assertion and Perspective. We also computed a similar difference score for the *raw counts* of PGMs coded as Assertion and Perspective—calculated for each subject as [# PGMs Perspective – # PGMs Assertion]—which we termed a P/A Difference score, to represent each individual's balance between the Overall Factors through the lens of total PGM counts. In Connors et al. (2014), we reported a range of values for both the P/A Difference score and the P/A Balance score in a sample of leaders, reflecting, as expected, individual differences across the subjects in terms of their cognitive motivational style as captured by MPA and the framework for decision making.

As discussed in Connors et al. (2014), inter-rater reliability was substantially higher for the patterning of PGMs within subject as detected by the P/A Balance Score (ICC = 0.89, CI = 0.77–0.95) as compared to the raw counts of PGMs (ICC = 0.41, CI = 0.02–0.69). A comparison of the 95% confidence intervals for these ICC coefficients reveals a lack of overlap, providing evidence that the difference is statistically significant.

Pattern Decoders

We reason that this comparison reveals a major point about observational methodology in general, and in particular the distinctive capture made in movement analysis. The strength of the MPA model is that it zeroes in on movement patterns indicative of embodied cognition—authentic, real-time actions that support and represent higher-order thought. The utility of the MPA model and coding system, and the skill of the expert MPA analyst, comes from training and experience in being highly attuned moment-to-moment to each individual's personal baseline and its relative balancing of the complementary decision-making processes. One could argue that in making these observations astutely the total raw counts are not as important as

deciphering the stable patterning demonstrated by the individual within the interview setting.

The analogy is a clinician who knows how to get exactly the right amount of data when interviewing a patient to make a diagnosis, as in the case of administering a semi-structured interview for a psychiatric disorder such as depression. While the interview yields quantitative information (e.g., length of time feeling depressed, ratings of symptom severity, symptom counts), there is a profound qualitative element in the sense of the clinician intuiting how best to arrive at the formulation required for that purpose—and in fact knowing just when it is that sufficient information has been acquired to finalize a diagnosis. The same holds for the MPA expert analyst, especially in terms of deciphering the inimitable way each individual displays signature phrases that replicate throughout a full observation period.

The broader point is that the micro coding of the PGMs represents a quantitative method that serves to guide the expert MPA analyst in deducing qualitative patterns that exhibit a macro understanding of an individual as a decision-maker. In short, the PGM Action Motivations also point to levels of Assertion and Perspective as well as sequencing—just how the decision maker is staging their cycle of decision making—whether for example they begin at Committing and then Intend and finally Attend or more locally Attend, Intend and Commit (see Connors et al., 2018a). This blending of the micro and macro and of the quantitative and qualitative is emblematic of the deep method of MPA analysis—the sensitivity to the patterning of movements within each individual is what expert analysts are trained to see— and thus the MPA analyst is adept at revealing an individual's management style.

DOES MPA SHOW EVIDENCE OF PREDICTIVE VALIDITY?

Establishing high inter-rater reliability of a coding system ensures that raters use it in a similar, consistent manner. This does not tell us if the coding system is performing as it is designed in order to predict future behavior. In the case of MPA, which is predicated on decoding embodied decision-making style, we would want to know if the reliable MPA profile is capable of offering prognostic value for how people will make future decisions at other times and in other settings. In order words, we would need to establish the predictive validity of MPA, which is the essence of establishing an empirical basis for using it to decode cognitive motivational decision-making style.

This brings us to the interdisciplinary framework necessary to evaluate the extent to which MPA offers predictive validity. We have aimed to build a bridge with decision-making science in order to identify ways of measuring what MPA was designed to predict, namely how people will stylistically go about making future decisions, and the corollary of being able to discriminate how a group of people will vary in this process. Here we outline steps that we have taken in our experimental work over the last 6 years that integrate constructs and measures from other disciplines to pursue this interdisciplinary aim (see Connors et al., 2013).

Measuring Individual Differences in Decision-Making Processes

A fundamental consideration was to work out the best way to measure individual differences in decision-making processes that could be recorded in real time and serve as the dependent variable to be predicted by MPA. A review of the prior literature at the time (see Connors et al., 2013) stimulated key considerations that guided our work. A primary recognition is that many paradigms used to directly study people as they engaged in decision making were highly constrained and as such would tend to diminish, rather than expose, differences in decision-making style (Connors et al., 2018a). We summed this up as follows:

" … many of the experimental methods used to study decision-making behavior can overwhelm or diminish the impact of individual differences—meaning that other designs need to be entertained. For example, very detailed instructions, strong manipulations within a paradigm, and highly restrictive forced choices (especially two-choice options) can dilute the role of personal characteristics in the experimental setting. It is critical that research methods be employed that can better simulate the real-world context of decision-making … " (Connors et al., 2013, p. 3).

We turned to a paradigm that has been used across different disciplines: the design and application of hypothetical decision-making scenarios that can be administered in a laboratory setting. A key consideration was that the interest was not in what decisions were made *per se* but in how individuals went about making them. Hypothetical scenarios could be utilized for this purpose, particularly if we attended to a fundamental design principle—subjects need to be granted some form of control over the information they could utilize to inform their decision making, and time constraints needed to be removed:

"To this end, we permitted subjects the freedom to control their own information search via the option of making requests for more information … as it is assumed that decision style would be influential in shaping this aspect of the process … decision style should be reflected in the strategies and motivations that guide information search (as some individuals would lean toward acquiring more versus less information before coming to a decision) as well as response time (as those who seek out more information would also spend more time before coming to a decision)" (Connors et al., 2013, p. 4).

The parameters referenced in the above quote—information search and response time—may be considered as indicators of the predecisional stage of decision making that may be especially resonant of individual differences in decision-making style when recorded while subjects engaged in hypothetical scenarios which offered them control over information and no time constraints. Importantly, we proposed such predecisional stages as articulated in the MPA model and captured by the coding system—again, the core is the manner in which individuals approach decision-making situations, particularly in terms of the priorities and cognitive motivations that define their stylistic propensities.

To generate outcome measures to gauge predictive validity, we created hypothetical decision-making tasks representing four domains (Financial, Health, Voting, and Strategy) drawn from prior work in political science (e.g., decision tree modeling) as well as behavioral research (see Connors et al., 2013). To facilitate expression of individual differences, subjects were provided options to request, one at a time, an additional piece of information to consider before they registered their decision. Subjects could either move on to make a final decision, or request another piece of information, in an iterative manner, at each step. In this way, the number of information requests (information search) could be tallied for each subject, and the amount of chronological time (decision time) spent before a decision was reached could be recorded.

Analytic Approaches and Findings

By crossing these two different methodologies across time—MPA as a baseline measure of the cognitive motivational decision style, and laboratory recordings of information search and decision time via the hypothetical scenarios—we had a platform to examine the predictive validity of MPA.

Our first benchmark paper revealed that the P/A Balance score was a robust predictor of individual differences in leaders' decision-making processes, as elicited in the decision-making scenarios. Specifically, a propensity for more Perspective was highly correlated with requesting more pieces of information and devoting more chronological time before coming to a decision (Connors et al., 2013). The correlations were in the "high" effect size range, suggesting robust prediction given the expectations in social science research.

We next examined the relative predictive values of the P/A Balance Score as compared to that offered by the P/A Difference score. In other words, we wanted to determine if the superior reliability of the patterning captured by the P/A Balance Score offered stronger and independent prediction of future decision-making processes as compared to raw count tallies of PGMs. This was indeed the case, as confirmed in stepwise regression models (Connors et al., 2014). The implication is that it is most informative to understand each individual's unique patterning of signature movements—how they balance their predilections for Assertions vs. Perspective relative to their own baseline of total PGMs—rather than absolute counts of PGMs as distributed when gaining prognostic insight from the MPA profile. The "balance" within an individual is the essence of decoding cognitive motivation as it applies to decision-making style.

By utilizing the strategy of creating balance scores using MPA data, we also showed that the individual differences in MPA that were most strongly predictive of the outcome measures could be localized within specific stages of the MPA model (Connors et al., 2015, 2018b). Initial work revealed that the way leaders differed in the way they balanced Assertion and Perspective during the 2nd Intending stage of the MPA model was especially predictive of individual differences in the indicators of predecisional processes measured during navigation of the decision-making scenarios. In particular, leaders who lean toward more Perspective via Evaluating (positioning oneself to size up and crystallize the options and set priorities) requested more pieces of information

and devoted more chronological time to the scenarios before coming to a decision, as compared to leaders who were more predisposed toward Assertion via Determining (devoting effort to build the resolve necessary to formulate a position and take a stand). A follow-up investigation utilized a larger sample size to not only replicate these findings, but also utilized factor analysis to show that individual differences in the balance between Assertion and Perspective most predictive of future decision-making processes were localized to the Intending and Committing stages in the MPA model (vs. the Attending stage), providing specificity in terms of how leaders differ in the cognitive motivations that drive the decision-making cycle.

The empirical work done to date complements the decades of application of MPA that has yielded ecological validity within organizations (Connors et al., 2018a). MPA is proving to deliver what is required of measures of decision-making style—significant prediction of how people, including leaders, differ in how they will navigate ambiguous and complex decision-making situations. Leaders across many types of organizations (e.g., corporate, public service, military, academia) face increasing challenges in the types of decisions to be made, the range of information to consider, and the relative stakes attached to their decisions. It has been posited that observational measures of decision-making style would offer prognostic power that transcends the typically moderate prediction provided by self-assessment instruments (Connors et al., 2016), and the findings to date on MPA are certainly in line with that assertion.

FUTURE DIRECTIONS

It is our hope that new phases of research will be launched that will broaden the interdisciplinary application and study of MPA. Movement resides at the intersection of thought and behavior, and we suggest we can discover deeper truths about what is meant by the term "embodied" as methods like MPA become incorporated into a range of disciplines. Here we focus on two potential future directions that we perceive to be of immediate interest.

Automated Coding

One question that can be raised is the extent to which observational methods such as MPA which focus on detection of complex movement patterns should be facilitated by automated coding. This is a complex issue which is of high relevance because there is a substantial literature devoted to the evolution of a number of automated methods designed to recognize and code nonverbal behavior, including human movement, particularly behavior indicative of psychological processes. Here we break out the issue in a number of ways to stimulate further thinking.

First, there are assumptions that automated coding is inherently superior to coding done by human observers, no matter how highly trained and skilled they may be. For example, there have been suggestions that nonverbal behavioral coding systems executed manually by human observers—including those designed to recognize and code a range of gestures—can suffer from low reliability and a substantial or complete lack of objectivity (Lauberg and Sloetjies, 2016; Mahmoud and

Robinson, 2016). While we do not underestimate the challenges involved in coding movement, we do not agree that human observers are fundamentally non-objective and machines are inherently unbiased. Every coding system is, by design, selective, in that specific behaviors are prioritized and coded, and others are ignored, whether the coding is done manually or by machine. Such selectivity needs to be rooted in theory and baseline study, and articulated to a sufficient degree that humans and/or machines can detect the full range of movements as specified in the coding system. Thus reliability becomes the metric of interest, and both humans and machines need to be evaluated empirically to determine the level of "objectivity" of a coding system. With respect to MPA, we defer to our prior section on inter-rater reliability as proof of principle that very high reliability of human coding can be achieved with sufficient standards of training of expert analysts. In other words, certified MPA analysts can achieve the type of objectivity that is desired in automated coding.

A complementary perspective on machine coding of nonverbal behavior, including gesture, should also be considered, as it flips the "bias" in terms of human vs. machine coding. It has been suggested that while much progress has been made in machine recognition of nonverbal indicators of cognitive states (Mahmoud et al., 2016), automated coding systems are often deficient in terms of predictive validity for a range of psychological functions (Lauberg and Sloetjies, 2016). This is a critically important point, in that it reinforces the complexity of detecting the meaning in human movement as it unfolds in real time, and the challenge of designing a machine to achieve the insight of highly trained human observers with decades of experience. We posit that the human element—the level of knowledge and insight offered by expert MPA analysts—will remain essential, especially as it would be what is transferred to a machine. We refer back to the essence of the MPA method, which requires the coding system to detect the signature patterns of an individual. What might prove to be prohibitively difficult for an automated system is generating the baseline—the qualitative understanding of the one-of-a-kind patterning of an individual's movement as it emerges across time and context. The experienced MPA analyst will be able to bring a clinician's rigorous observational experience and deep insight to recognize if a sufficient representative patterning of PGMs has been achieved to permit a valid MPA analysis of a given individual. There are times, for example, when an individual will not produce enough PGMs to permit the creation of an MPA profile (Lamb, 2012), and which may prove impossible to determine by an automated coding system. In addition, when applying MPA, the expert MPA analyst offers cogent interpretation and coaching in tune with the generated profile (Connors et al., 2018a), and that is an essential human skill and deliverable of the MPA process which cannot be automated. As we have shown, this level of insight translates into high predictive validity of how leaders differ in their navigation of decision-making scenarios, and any translation of a complex coding system such as MPA would need to be shown to achieve the same result.

Taken together, we suggest that it is not essential for coding systems such as MPA to be executed by machines in order to achieve the objectivity, reliability, and predictive validity we would demand. That said, the idea of moving toward automated coding of MPA should not be dismissed and in fact should be entertained in future research. The reality is that meaningful observational systems of complex behaviors such as movement patterns requires much time and effort on the part of highly trained human coders (Velloso et al., 2013; Schreer et al., 2014; Mahmoud and Robinson, 2016). It could be advantageous to facilitate the application of MPA to explore methods for integrating automated methods that might be implemented without eliminating the essential expert perspective; such efforts could, for example, permit either more rapid coding of an MPA interview along with facilitating broader application to larger numbers of individuals. While we posit that the complexity of detecting PGMs—highly integrated movement patterns that require perception of the whole body in motion—will prove to be a substantial challenge for developers of automated coding systems, there are studies which have shown initial levels of success at representing the body movements of a samba dancer (Chavoshi et al., 2015), along with application of LMA to detect hand movements (Lourens et al., 2010) and segmentation of a repertoire of motions (Bouchard and Badler, 2007).

One area that could be impactful would be to utilize expert human coding to scan and review videotape of a subject and isolate particularly meaningful segments that could, in principle, eventually be coded by machine. An experienced MPA analyst is skilled in detecting if an individual is revealing notable replicable patterning—particularly with respect to "peak performance," a patterning akin to an athlete observed with all cylinders firing—which affirms the patterning and finalization of an MPA profile. Future research could work through methodological designs to determine if automated coding of such peak performances would yield the same level of reliability and validity achieved by MPA analysts. If this proves to be the case, this could facilitate more rapid coding of MPA, which could support time-sensitive demands as well as broader application of the method by reducing coding time.

Neuroscience and Different Elements of Decision-Making Style

A second core area we envision to be of high importance would be future collaborative efforts that incorporate neuroscience perspectives to expand our understanding of decision making as a fundamentally embodied human behavior. Many advances have been made in deciphering complex neural networks that underlie the cognitive (or executive) functions that go into the decision making process (Rosenbloom et al., 2012). Insight is also accruing on how people differ in the way they make decisions using a range of neuroscience methods (see Connors et al., 2016). For example, Talukdar et al. (2018) focused on seven intrinsic connectivity networks (ICNs) that showed promise as a framework for understanding how individuals differ across multiple components (executive, perceptual, social) of decision making.

There has also been attention given to neural models of embodied cognition. However, these have tended to focus more

on neural activation of motor systems in relation to cognition—such as via study of the mirror neuron system (Keysers et al., 2018) and how sensorimotor areas are involved in the processing of action words (Gallese and Cuccio, 2018; Horoufchin et al., 2018)—rather than take on the question of how movement integrates with cognition at the neural level. That said, there have been advances in neuroscience, at the level of findings and conceptual thinking, which certainly converge with the notion of moving toward a neuroscience of embodied decision making.

Perhaps most prominently, current conceptions of the functional capacity of the cerebellum—the "motor control" center of the brain—have uncovered the multiple functions mediated by this structure, which include affective/emotional processes along with perceptual/cognitive operations, echoing the Latin origin of the word cerebellum ("little brain"). Of particular note is that authoritative consensus statements have been made by leading neuroscientists recognizing that the cerebellum plays a much larger role than has been assumed in both perception (Baumann et al., 2015) and cognition (Koziol et al., 2014). As researchers focus more extensively on the cerebellum—both in terms of the range and complexities of functions it subsumes, and the structural and functional connections with subcortical and cortical networks—much deeper insight into how people are wired to be embodied decision makers may be achieved.

A primary impetus for the expanded conception of the functions subsumed by the cerebellum was clinical observation of cognitive deficits in patients suffering from a range of neurological disorders that, in principal, are strictly movement disorders. It became clear, based on a critical mass of clinical studies, that pathology of movement areas in the brain can compromise cognitive capacities, providing proof of principle that movement and cognition are not just functionally connected but also integrated at the neural level. A similar viewpoint is emerging in research on Parkinson's disease, particularly in terms of decision making. Perugini et al. (2018) noted that decision making is often impaired in people with Parkinson's disease, including disruption of the ability to gather a sufficient amount of information necessary to make a decision. They suggest that it will be fruitful to apply a decision-making framework to gain a better understanding of the intersection of motor and cognitive difficulties—including emphasis on the neural circuits that subsume perceptual decision-making and integration with memory—that manifest in the trajectory of Parkinson's disease.

This idea of integration of multiple sensory systems in the brain may hold particular promise for catalyzing our capacity to unravel the profundity offered by embracing the embodied nature of cognition, including decision making. Pasqualotto et al. (2015) emphasize that the recognition of the fundamental nature of multisensory integration involves the interconnections between the "brain, body, and world" which serves as an elegant way of encapsulating why decision making should be

conceptualized as an embodied form of behavior. A glimpse into the future is offered by Ryu and Torres (2018), who presented both a theoretical and statistical framework to guide explorations of the integration of biophysical activity with cognitive activity when engaging in decision making. Their novel work revealed, for example, how increases in cognitive load during decision making led to dynamic changes in multiple physiological systems, including hand movements along with heart signals. We anticipate seeing more research that uncovers such elegant real-time measurement of integrated nervous system activity (body and brain) that mobilizes during decision making.

A related area to be considered in future neuroscience work is to incorporate movement-based coding systems such as MPA, along with more traditional measures of decision-making style, which are typically based on self-report.

Indeed, there is little research on the relative distinctions between how people perceive themselves as decision makers, and their signature cognitive motivations detected by MPA that may elude self-reflection and self-perception. As consideration of both of these elements would feed into achieving a more complete understanding of the decision maker, we suggest that eventual convergence with neuroscience methods can begin to tease apart and map out the different components of decision-making style, and of the individual decision maker. It is likely that there are a number of neural networks that differentiate how we see ourselves as decision makers, and the motivations that drive our priorities as a decision maker.

CONCLUSION

Our overall conclusion is that we are witnessing a new era of understanding not just about decision making, but about decision makers—and our traction is greatly aided by appreciating and assessing that people are embodied creatures who rely on movement to drive and navigate their own priorities as decision makers. We qualitatively self-navigate the quantitative mechanical world of distance, force, time and flow via movement, which serves essential functions in integrating the interplay between our internal states and our environment, and our sensorimotor integration that drives both thought and action. Future research that can sharpen our ability to acquire deeper insight into embodied decision making and decision makers would offer much in the way of optimizing human performance across many domains as well as guide therapeutic efforts that operate through a deep understanding of movement (Connors et al., 2018a).

AUTHOR CONTRIBUTIONS

BC and RR equally contributed to conceptualizing the paper, reviewing references and materials, and writing and editing the submitted manuscript.

REFERENCES

Bakeman, R., and Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences.* Cambridge: Cambridge University Press.

Baumann, O., Borra, R. J., Bower, J. M., Cullen, K. E., Habas, C., Ivry, R. B., et al. (2015). Consensus paper: the role of the cerebellum in perceptual processes. *Cerebellum* 14, 197–220. doi: 10.1007/s12311-014-0627-7

Bouchard, D., and Badler, N. (2007). "Semantic segmentation of motion capture using Laban movement analysis," in *Intelligent Virtual Agents, IVA 2007*, Lecture Notes in Computer Science, Vol. 4722, eds C. Pelachaud, J. C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé (Berlin; Heidelberg: Springer).

Chavoshi, S. H., De Baets, B., Neutens, T., Delafontaine, M., De Tre, G., and Van de Weghe, N. (2015). Movement pattern analysis based on sequence signatures. *ISPRS Int. J. Geo Inform.* 4, 1605–1626. doi: 10.3390/ijgi4031605

Connors, B. L., Moore, C.-L., Rende, R., and Colton, T. J. (2018a). "Movement Pattern Analysis (MPA): decoding individual differences in embodied decision making," in *The SAGE Handbook of Personality and Individual Differences,* eds V. Zeigler-Hill and T. Shackelford (New York, NY: SAGE), 257–277.

Connors, B. L., Rende, R., and Colton, T. J. (2013). Predicting individual differences in decision-making process from signature movement styles: an illustrative study of leaders. *Front. Psychol.* 4:658. doi: 10.3389/fpsyg.2013.00658

Connors, B. L., Rende, R., and Colton, T. J. (2014). Inter-rater reliability for Movement Pattern Analysis (MPA): measuring patterning of behaviors versus discrete behavior counts as indicators of decision-making style. *Front. Psychol.* 5:605. doi: 10.3389/fpsyg.2014.00605

Connors, B. L., Rende, R., and Colton, T. J. (2015). Decision-making style in leaders: uncovering cognitive motivation using signature movement patterns. *Int. J. Psychol. Stud.* 7, 105–112. doi: 10.5539/ijps.v7n2p105

Connors, B. L., Rende, R., and Colton, T. J. (2016). Beyond self-report: emerging methods for capturing individual differences in decision-making process. *Front. Psychol.* 7:312. doi: 10.3389/fpsyg.2016.00312

Connors, B. L., Rende, R., and Colton, T. J. (2018b). Staging of decision-making style in leaders: decoding individual differences in cognitive motivations using Movement Pattern Analysis. *Funct. Neurol. Rehabil. Ergon.* 7, 14–19.

Dishion, T. J., and Snyder, J. (2004). An introduction to the special issue on advances in process and dynamic system analysis of social interaction and the development of antisocial behavior. *J. Abnorm. Child Psychol.* 32, 575–578. doi: 10.1023/B:JACP.0000047317.96104.ca

Gallese, V., and Cuccio, V. (2018). The neural exploitation hypothesis and its implications for an embodied approach to language and cognition: insights from the study of action verbs processing and motor disorders in Parkinson's disease. *Cortex* 100, 215–225. doi: 10.1016/j.cortex.2018.01.010

Girard, J. M., and Cohn, J. F. (2016). A primer on observational measurement. *Assessment* 23, 404–413. doi: 10.1177/107319.

Horoufchin, H., Bzdok, D., Buccino, G., Borghi, A. M., and Binkofski, F. (2018). Action and object words are differentially anchored in the sensory motor system – A perspective on cognitive embodiment. *Sci. Rep.* 8:6583. doi: 10.1038/s41598-018-24475-z

Keysers, C., Paracampo, R., and Gazzola, V. (2018). What neuromodulation and lesion studies tell us about the function of the mirror neuron system and embodied cognition. *Curr. Opin. Psychol.* 24, 35–40. doi: 10.1016/j.copsyc.2018.04.001

Koziol, L. F., Budding, D., Andreasen, N., D'Arrigo, S., Bulgheroni, S., Imamizu, H., et al. (2014). Consensus paper: the cerebellum's role in movement and cognition. *Cerebellum* 13, 151–177. doi: 10.1007/s12311-013-0511-x

Laban, R. (1928). *Schrifttanz: Methodik, Orthographie, Eriauterungen.* Vienna: Universal Edition.

Laban, R. (1950). *The Mastery of Movement on the Stage.* London: MacDonald & Evans.

Laban, R., and Lawrence, F. C. (1947). *Effort.* London: MacDonald & Evans.

Lamb, W. (2012). *A Framework for Understanding Movement: My Seven Creative Concepts.* London: Brechin Books.

Lauberg, H., and Sloetjes, H. (2016). The revised NEUROGES-ELAN system: an objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture. *Behav. Res. Methods* 48, 973–993. doi: 10.3758/s13428-015-0622-z

Lepora, N. F., and Pezzulo, G. (2015). Embodied choice: how action influences perceptual decision making. *PLoS Comput. Biol.* 11:e1004110. doi: 10.1371/journal.pcbi.1004110

Lourens, T., van Berkel, R., and Brakova, E. (2010). Communicating emotions and mental states to robots in a real time parallel framework using Laban movement analysis. *Robot. Auton. Syst.* 58, 1256–1265. doi: 10.1016/j.robot.2010.08.006

Mahmoud, M., and Robinson, P. (2016). "Towards automatic analysis of gestures and body expressions in depression," in *Pervasive Health'16: Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Cancún), 276–277.

Mahmoud, M., Baltrusaitis, T., and Robinson, P. (2016). Automatic analysis of naturalistic hand-over-face gestures. *ACM Trans. Interact. Intell. Syst.* 6:19. doi: 10.1145/2946796

Moore, C. L. (2005). *Movement and Making Decisions: The Body-Mind Connection In The Workplace.* New York, NY: Rosen Publishing Group.

Pasqualotto, A., Dumitru, M. L., and Myachykov, A. (2015). Editorial: multisensory integration: brain, body, and world. *Front. Psychol.* 6:2046. doi: 10.3389/fpsyg.2015.02046

Perugini, A., Ditterich, J., Shaikh, A. G., Knowlton, B. J., and Basso, M. A. (2018). Paradoxical decision-making: a framework for understanding cognition in Parkinson's disease. *Trends Neurosci.* doi: 10.1016/j.tins.2018.04.006. [Epub ahead of print].

Pietrazak, T., Lohr, C., Jahn, B., and Hauke, G. (2018). Embodied cognition and the direct induction of affect as a complement to cognitive behavioural therapy. *Behav. Sci.* 8:e29. doi: 10.3390/bs8030029

Rende, R., Conti, D. V., Gilman, S. E., and Slomkowski, C. (2009). Incorporating social context into genetic studies of nicotine dependence. *NCI Tobacco Control Monograph 22* (NIH). Bethesda.

Rosenbloom, M. H., Schmahmann, J. D., and Price, B. H. (2012). The functional neuroanatomy of decision making. *J. Neuropsychiatry Clin. Neurosci.* 24, 266–277. doi: 10.1176/appi.neuropsych.11060139

Ryu, J., and Torres, E. B. (2018). Characterization of sensory-motor behavior under cognitive load using a new statistical platform for studies of embodied cognition. *Front. Hum. Neurosci.* 12:116. doi: 10.3389/fnhum.2018.00116

Schreer, O., Masneir, S., Skomroch, H., and Lausdberg, H. (2014). "Coding hand movement behavior and gesture with NEUROGES supported by automatic video analysis," in *9th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2014)* (Wageningen).

Slomkowski, C., Conger, K. J., Rende, R., Heylen, E., Little, W. M., Shebloski, B., et al. (2009). Sibling contagion for drinking in adolescence: a micro process framework. *Eur. J. Dev. Psychol.* 3, 161–171.

Talukdar, T., Román, F. J., Operskalski, J. T., Zwilling, C. E., and Barbey, A. K. (2018). Individual differences in decision making competence revealed by multivariate fMRI. *Hum. Brain Mapp.* 39, 2664–2672. doi: 10.1002/hbm.24032

Velloso, E., Bulling, A., and Gellersen, H. (2013). "AutoBAP: automatic coding of body action and posture units from wearable sensors," in *Humane Association Conference on Affective Computing and Intelligent Interaction: IEEE Computer Society* (Los Alamitos, CA), 135–140.

Winter, D. D. (1992). "Body movement and cognitive style: validation of action profiling," in *The Body-Mind Connection in Human Movement Analysis*, ed S. Lomm (Keene, NH: Antioch New England Graduate School).

Winter, D. D., Widell, C., Truitt, G., and George-Falvy, J. (1989). Empirical studies of posture-gesture mergers. *J. Nonverbal Behav.* 13, 207–223. doi: 10.1007/BF00990294

Investigating the Correspondence of Clinical Diagnostic Grouping With Underlying Neurobiological and Phenotypic Clusters Using Unsupervised Machine Learning

Xinyu Zhao [1,2], D. Rangaprakash [1,3], Bowen Yuan [1], Thomas S. Denney Jr [1,4,5,6], Jeffrey S. Katz [1,4,5,6], Michael N. Dretsch [7,8] and Gopikrishna Deshpande [1,4,5,6,9*]

[1] Department of Electrical and Computer Engineering, AU MRI Research Center, Auburn University, Auburn, AL, United States, [2] Quora, Inc., Mountain View, CA, United States, [3] Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA, United States, [4] Department of Psychology, Auburn University, Auburn, AL, United States, [5] Alabama Advanced Imaging Consortium, Auburn University, University of Alabama at Birmingham, Birmingham, AL, United States, [6] Center for Neuroscience, Auburn University, Auburn, AL, United States, [7] Human Dimension Division, HQ TRADOC, Fort Eustis, VA, United States, [8] U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL, United States, [9] Center for Health Ecology and Equity Research, Auburn University, Auburn, AL, United States

Many brain-based disorders are traditionally diagnosed based on clinical interviews and behavioral assessments, which are recognized to be largely imperfect. Therefore, it is necessary to establish neuroimaging-based biomarkers to improve diagnostic precision. Resting-state functional magnetic resonance imaging (rs-fMRI) is a promising technique for the characterization and classification of varying disorders. However, most of these classification methods are supervised, i.e., they require *a priori* clinical labels to guide classification. In this study, we adopted various unsupervised clustering methods using static and dynamic rs-fMRI connectivity measures to investigate whether the clinical diagnostic grouping of different disorders is grounded in underlying neurobiological and phenotypic clusters. In order to do so, we derived a general analysis pipeline for identifying different brain-based disorders using genetic algorithm-based feature selection, and unsupervised clustering methods on four different datasets; three of them—ADNI, ADHD-200, and ABIDE—which are publicly available, and a fourth one—PTSD and PCS—which was acquired in-house. Using these datasets, the effectiveness of the proposed pipeline was verified on different disorders: Attention Deficit Hyperactivity Disorder (ADHD), Alzheimer's Disease (AD), Autism Spectrum Disorder (ASD), Post-Traumatic Stress Disorder (PTSD), and Post-Concussion Syndrome (PCS). For ADHD and AD, highest similarity was achieved between connectivity and phenotypic clusters, whereas for ASD and PTSD/PCS, highest similarity was achieved between connectivity and clinical diagnostic clusters. For multi-site data (ABIDE and ADHD-200), we report site-specific results. We also reported the effect of elimination of outlier subjects for all four

datasets. Overall, our results suggest that neurobiological and phenotypic biomarkers could potentially be used as an aid by the clinician, in additional to currently available clinical diagnostic standards, to improve diagnostic precision. Data and source code used in this work is publicly available at https://github.com/xinyuzhao/identification-of-brain-based-disorders.git.

INTRODUCTION

A neuropsychiatric disorder is a brain-based dysfunctional condition associated with impairments of affect, cognition and behavior. Many factors contribute to these disorders, e.g., genes, family history, substance abuse, traumatic brain injury, life experience, environmental conditions etc. Conventional diagnosis primarily consists of clinical interviews and standardized psychometric testing, which are recognized to be largely imperfect [1–3]. Because neuropsychiatric pathologies are complex, which can lead to inconsistencies between clinicians' diagnoses, there is increasing interest in identifying non-invasive neuroimaging biomarkers. The most commonly used approach for achieving this is by employing *supervised* learning models such as support vector machines [4, 5], artificial neural networks [6], and decision trees [7], wherein the model learns the associations between patterns in the data and diagnostic labels using a training data set. This model can then be tested on an independent validation data set. However, the problem with such an approach is that the model itself is based on clinical labels, and hence, it cannot be used to uncover novel structures and groupings from the data. The problem can be addressed by employing *unsupervised* models. *Unsupervised* models have been traditionally used to uncover clusters of subjects with similar patterns of imaging data, with applications in identifying disease clusters [8, 9] as well as sub-clusters [10] within a disease. Most of these studies use k-means clustering [11, 12]. However, it is besieged with methodological issues such as *a priori* choice of clusters needed in k-means, and the large dimensionality of imaging data necessitates some type of dimensionality reduction for clustering to work as intended. Problematically, this last step is either not carried out [13], or carried out by preselecting features not from the structure in the data, but by some external considerations such as previous findings in a given disorder [14, 15]. Such approaches rob the method of its advantages of being truly data-driven, in that the clusters obtained from imaging data are seldom compared to data obtained from clinical diagnostic criteria and related behavioral phenotypes. This is important because disease clusters obtained from any method, be it imaging or another diagnostic tool, should be linked with the behavioral phenotype. In this study, we address the above shortcomings using resting state functional magnetic resonance imaging (rs-fMRI) data obtained from five different neuropsychiatric disorders: Attention Deficit Hyperactivity Disorder (ADHD), Alzheimer's Disease (AD), Autism Spectrum Disorder (ASD), Posttraumatic Stress

Disorder (PTSD) and Post-Concussion Syndrome (PCS). We provide a brief introduction to these disorders in the following paragraphs.

Attention Deficit Hyperactivity Disorder

ADHD is a psychiatric disorder characterized by impulsiveness, inattention, and hyperactivity. This condition affects about 5% of children and adolescents worldwide [16]. Symptoms include difficulty staying focused and paying attention, difficulty controlling behavior, and hyperactivity. ADHD has three subtypes: ADHD hyperactive-impulsive (ADHD-H), ADHD inattentive (ADHD-I), and ADHD combined hyperactive-impulsive and inattentive subtype (ADHD-C). Because symptoms vary from person to person, ADHD can be difficult to identify. In addition, there has been a debate that ADHD is over-diagnosed in children and adolescents by current clinical criterion [17].

Alzheimer's Disease

AD is the most commonly diagnosed type of dementia in elderly patients [18], which is characterized by memory dysfunction, cognitive decline, etc. Before the onset of dementia, patients may develop an intermediate stage of dysfunction known as mild cognitive impairment (MCI). Patients with MCI have a higher risk of progressing to AD [19]. According to results from the Honolulu-Asia Aging Study [20], as many as one-third of all Alzheimer's diagnoses may actually be false positives. In addition, the diagnostic boundary between AD and MCI is not well established.

Autism Spectrum Disorder

ASD is a pervasive developmental disorder clinically characterized by social and communication impairments as well as restricted interests and repetitive behaviors [1]. While the boundaries between ASD, its comorbidities, and neurotypicals with sub-clinical ASD-like traits are blurred, several diagnostic subcategories within ASD were defined: autism, Asperger's disorder, and pervasive developmental disorder-not otherwise specified (PDD-NOS). It has been often argued that the Asperger's disorder criteria is problematic [21, 22]. In the latest DSM-V classification, Asperger's and PDD-NOS were eliminated, in favor of the so called "dimensional assessment" of the autism spectrum [23]. This highlights the confusion in the field due to lack of objective biomarkers based on underlying neurobiology.

Post-traumatic Stress Disorder

PTSD is a disabling condition in individuals exposed to a traumatic event, such as war, violent crime, and motor vehicle accidents [24]. PTSD is characterized by intrusive avoidance, hypervigilance, hyperarousal and alterations in cognition and mood [25]. PTSD is found to be associated with aberrant functioning of the amygdala, hippocampus, insula, and regions of the prefrontal cortex such as the ventromedial PFC [26–28]. Although cognitive decrements are associated with PTSD, there is evidence that they are mediated by comorbid symptoms of the disorder (e.g., depression and anxiety) [29].

Post-concussion Syndrome

PCS emerges in a subset of individuals who sustain a mild traumatic brain injury (mTBI) or concussion. It includes a constellation of prolonged symptoms, which persist several months after the mTBI. Symptoms can be categorized as vestibular, cognitive, affective, and somatosensory [30–32]. In military service members, diagnosis can be more complex since it has high co-morbidity with PTSD, along with homogeneity of symptomatology between the two disorders [33, 34].

In summary, although the neuropsychiatric disorders delineated have well established diagnostic criteria, overlapping symptoms with other neuropsychiatric disorders are common, and in some cases manifest as comorbidities. The neural circuits implicated in a given disorder are also often overlapping with that of other disorders. In addition, conventional diagnostic categories often do not adequately capture the spectrum of symptoms and impairments ranging from mild to severe. Further, categorization of subgroups within several disorders have yet to be fully characterized. Thus, neuroimaging-based diagnostic classification and biomarkers can improve our understanding of subgroups within a specific neuropsychiatric disorder and eventually improve diagnostic precision.

This approach has indeed been promoted actively by the National Institute of Mental Health (NIMH) in the United States by the publication of "*Research Domain Criteria*" (RDoC, http://www.nimh.nih.gov/research-priorities/rdoc/nimh-research-domain-criteria-rdoc.shtml). RDoC is agnostic to present disorder categories. Its intent is the generation of disorder classifications in a data-driven fashion. The "*core unit of analysis*" as advanced by RDoC is the "*measurements of particular circuits as studied by neuroimaging techniques.*" In keeping with this ideology, a recent report demonstrated how data-driven definition of diagnostic groups in psychiatric spectrum disorders could identify new groups that had better mapping onto behavioral clusters [35, 36]. Our approach in this work is inspired by these recent developments.

Resting-state functional magnetic resonance imaging (rs-fMRI) is a promising tool for studying neuropsychiatric disorders [37–42]. It measures spontaneous fluctuations in the blood oxygen level-dependent (BOLD) signal, while the participants do not perform any explicit task [43, 44]. Considering machine learning applied to rs-fMRI, a common

methodology is to apply supervised classification methods on functional connectivity (FC) features obtained from rs-fMRI to identify brain-based disorders with FC aberrations. For example, some studies [6] used support vector machine (SVM) and artificial neural network (ANN) on different brain connectivity measures to identify ADHD. Khazaee et al. [3] combined a graph theoretical approach with SVM to classify patients with AD and MCI from healthy individuals. Plitt et al. [2] applied different classification methods, e.g., K-nearest neighbor (KNN), linear support vector machines (L-SVM), Gaussian kernel support vector machines (rbf-SVM) and L1-regularized logistic regression on rs-fMRI connectivity features to establish biomarkers for Autism spectrum disorders (ASD). However, the supervised machine learning methods used in such studies requires *a priori* clinical diagnoses to guide classification. In addition, most studies only target classifying one specific illness.

In this work, we attempt to address the aforementioned four challenges in *supervised* models by deriving a general analysis pipeline for identifying different neuropsychiatric disorders using *unsupervised* clustering methods. There have been very few studies using unsupervised models on rs-fMRI data to identify different neuropsychiatric disorders [1]. The main idea of unsupervised learning or clustering is to group subjects in such a way that those in the same group are more similar to each other than to those in other groups. Three clustering methods were specifically chosen: hierarchical clustering [45], ordering points to identify the clustering structure (OPTICS) [46], and density peak clustering (DPC) [47], since these methods do not require *a priori* specification of the number of clusters. The commonly used k-means clustering [48, 49] was not considered in this study due to the uncertainty of the number of clusters and sensitivity to outliers. Since clustering accuracy is often lower in high dimensional feature space, feature selection methods were applied. Most existing feature selection algorithms in the machine learning literature focus on heuristic search such as sequential forward searching (SFS) [50], non-linear optimization [51], genetic algorithm (GA) [50], etc. Bradley et al. proposed a non-linear optimization using a non-linear kernel support vector machine. Although this method provides high accuracy, it can only be used in the supervised learning context. SFS was proposed based on a greedy algorithm, which follows the problem-solving heuristic of making the locally optimal decision at each step. Similar to SFS, here we propose a sequential feature ranking (SFR) method by applying ANOVA testing among different groups (e.g., control group, disease subgroups) and then sequentially selecting features from the original dataset based on the *p*-value of each feature. Although SFS and SFR can be applied in unsupervised learning, they do not guarantee an optimal solution. Therefore, we propose GA as a robust feature selection method for unsupervised learning approaches for the identification of disease clusters from FC features by maximizing the similarity between connectivity and clinical diagnosis, as well as between connectivity and behavioral phenotypes, respectively. The identified clusters were then compared with those obtained from clinical diagnostic criteria and behavioral phenotypes in

FIGURE 1 | Illustration of proposed pipeline for identifying different brain-based disorders using unsupervised clustering methods. The main pipeline is depicted in purple color along with two supplementary analyses, site-specific analysis in salmon color and outlier subject elimination in green color. SFC, static functional connectivity; vDFC, variance of dynamic functional connectivity; SEC, static effective connectivity; vDEC, variance of dynamic effective connectivity; HRF, hemodynamic response functional; GA, genetic algorithm; DPC, density peak clustering; OPTICS, ordering points to identify the clustering structure: clustering algorithm; CM, co-association matrix.

order to investigate their similarity with disease clusters identified from rs-fMRI connectivity.

MATERIALS AND METHODS

In this work, a general pipeline has been derived (**Figure 1**) for identifying different brain-based disorders using unsupervised clustering methods. In addition, several supplementary analyses have been performed, e.g., site-specific analysis for multi-site data, elimination of outlier subjects, and enrichment analysis. The details of each step in the pipeline are described next. ADHD, AD and ASD data were obtained from publicly available databases. Details regarding ethical approvals for those data can be obtained from the links provided below. The PTSD study was carried out in-house, in accordance with the recommendations of Auburn University Institutional Review Board (IRB) and the Headquarters U.S. Army Medical Research and Material Command, IRB (HQ USAMRMC IRB) with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The

protocol was approved by the Auburn University IRB and the HQ USAMRMC IRB.

Participants and Non-imaging Measures
ADHD
Four hundred and eighty-seven subjects with complete phenotypic data were selected from the ADHD-200 dataset (http://fcon_1000.projects.nitrc.org/indi/adhd200/), which included 272 healthy controls (HC), 118 subjects with ADHD-C, and 97 subjects with ADHD-I. The number of subjects with ADHD-H were too small ($n < 10$), therefore ADHD-H was not considered in this work. The subjects were scanned at one of these three different sites: Peking University, Kennedy Krieger Institute (KKI), and New York University Child Study Center (NYU).

Subjects scanned at Peking University with diagnosis of ADHD were initially identified using the Computerized Diagnostic Interview Schedule IV [C-DIS-IV] [52]. All participants (ADHD and HC) were evaluated with the Schedule of Affective Disorders and Schizophrenia for Children—Present and Lifetime Version [KSADS-PL] [53], with one parent for

TABLE 1 | Phenotypic variables selected by GA with different clustering methods (ADHD).

Phenotypic/genetic variables	Selected variables		
	DPC	OPTICS	Hierarchical
ADHD index score	✓		
Inattentive score	✓	✓	✓
Hyper/impulsive score	✓	✓	✓
VIQ		✓	✓
PIQ			✓
FIQ	✓		✓

TABLE 2 | Phenotypic variables selected by GA with different clustering methods (AD).

Phenotypic/genetic variables	Selected variables		
	DPC	OPTICS	Hierarchical
APOE A1 and A2		✓	✓
NPI			
GDS			
MMSE	✓	✓	✓
CDR	✓	✓	✓
FAQ	✓	✓	✓

the establishment of the diagnosis for study inclusion. The ADHD Rating Scale [ADHD-RS-IV] [54, 55] was employed to provide dimensional measures of ADHD symptoms. Intelligence was evaluated with the Wechsler Intelligence Scale for Chinese Children-Revised [WISCC-R] [56].

In the KKI sample, psychiatric diagnoses were based on evaluations with the Diagnostic Interview for Children and Adolescents, Fourth Edition [DICA-IV] [57], a structured parent interview based on DSM-IV criteria; the Conners' Parent Rating Scale-Revised, Long Form [CPRS-R] [58], and ADHD-RS-IV. Intelligence was evaluated with the Wechsler Intelligence Scale for Children-Fourth Edition [WISC-IV] [59] and academic achievement was assessed with the Wechsler Individual Achievement Test-II [60].

In the NYU sample, psychiatric diagnoses were based on evaluations with KSADS-PL, administered to parents and children and CPRS-R. Intelligence was evaluated with the Wechsler Abbreviated Scale of Intelligence [WASI] [61].

Six Phenotypic variables were measured for all sites (**Table 1**), i.e., three ADHD measures including ADHD index score, Inattentive score, and Hyper/Impulsive score [62], and three IQ measures including Verbal IQ [VIQ], Performance IQ [PIQ], and Full Scale IQ [FIQ] [63].

AD

Rs-fMRI data from the Alzheimer's disease neuroimaging initiative (ADNI) database (http://adni.loni.usc.edu/) was utilized in this study. The sample consisted of subjects with three progressive stages of cognitive impairment—early MCI [EMCI] ($n = 23$), late MCI [LMCI] ($n = 29$), and AD ($n = 13$)—along with matched HC ($n = 31$).

The patients with AD had a Mini-Mental State Examination [MMSE] [64] score of 14–26, a Clinical Dementia Rating [CDR] [65] of 0.5 or 1.0 and met the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association [NINCDS/ADRDA] criteria [66] for probable AD. The patients with MCI had MMSE scores between 24 and 30, a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living and an absence of dementia [3].

Eight phenotypic variables, i.e., neuropsychiatric inventory [NPI] score [67], geriatric depression scale [GDS] [65], MMSE, CDR and functional assessment questionnaire [FAQ] [68], and one genetic variable i.e., apolipoprotein [APOE] A1 and A2 genotypes [69], were measured (**Table 2**). Except for the AD dataset, all other three datasets (ADHD, ASD, and PTSD) only have phenotypic variables. Thus, we just refer to these variables as phenotypic variables henceforth.

ASD

Four hundred and fifty-four subjects with complete phenotypic data were selected from the Autism Brain Imaging Data Exchange (ABIDE) database (http://fcon_1000.projects.nitrc.org/indi/abide/index.html). The sample consisted of 256 HC, 166 subjects with autism, and 32 subjects with Asperger's. Including PDD-NOS and "Asperger's or PDD-NOS" would have made the whole dataset more unbalanced, therefore these two subgroups were not considered in this study. Each subject was scanned at one of the following seven different sites: California Institute of Technology (Caltech), Carnegie Mellon University (CMU), NYU Langone Medical Center (NYU), University of Pittsburgh School of Medicine (Pitt), San Diego State University (SDSU), Trinity Center for Health Sciences (Trinity), and University of California Los Angeles (UCLA).

For most of the sites, diagnosis of ASD was consistent with the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision [DSM-IV-TR] criteria [70], and classification of either autism or Asperger's was made by a clinician based on the Autism Diagnostic Observation Schedule [ADOS] [71] and Autism Diagnostic interview-Revised [ADI-R] [72]. HC subjects were screened through a self-report history questionnaire to rule out other disorders, such as ASD, ADHD, or Tourette's Disorder.

Ten phenotypic variables were measured (**Table 3**) at all sites including three IQ measures, i.e., FIQ, VIQ, PIQ, four ADI_R measures, i.e., Reciprocal Social Interaction Subscore [ADI_R_SOCIAL], Abnormalities in Communication Subscore [ADI_R_VERBAL], Restricted, Repetitive, and Stereotyped Patterns of Behavior Subscore [ADI_RRB], Abnormality of Development Evident at or Before 36 Months Subscore [ADI_R_ONSET], and three ADOS measures, i.e., Classic Total ADOS Score [ADOS_TOTAL], Communication Total Subscore

TABLE 3 | Phenotypic variables selected by GA with different clustering methods (ASD).

Phenotypic/genetic variables	Selected variables		
	DPC	OPTICS	Hierarchical
FIQ	✓	✓	✓
VIQ	✓		
PIQ	✓		
ADI_R_SOCIAL	✓	✓	
ADI_R_VERBAL		✓	✓
ADI_RRB		✓	
ADI_R_ONSET			
ADOS_TOTAL			✓
ADOS_COMM		✓	✓
ADOS_SOCIAL	✓		✓

of the Classic ADOS [ADOS_COMM], and Social Total Subscore of the Classic ADOS [ADOS_SOCIAL].

PTSD and PCS

Eighty-seven active-duty male U.S. Army Soldiers (selected from Fort Benning, GA, USA and Fort Rucker, AL, USA) voluntarily participated in the current study. All of subjects had combat experience in Iraq (Operation Iraqi Freedom [OIF]) and/or Afghanistan (Operation Enduring Freedom [OEF]). Each subject was evaluated using three factors: (1) symptom severity in PTSD measured with "PTSD Checklist-5" [PCL5] score [73], (2) symptom severity in PCS measured with "Neurobehavioral Symptom Inventory" [NSI] score [74], and (3) medical history. Based on these factors, (i) 17 subjects were grouped as PTSD with no history of mTBI in the last 5 years, a total score ≥38 on PCL5 and <26 on NSI, (ii) 42 subjects were grouped as the comorbid PCS+PTSD with a history of medically documented mTBI, post-concussive symptoms, and scores ≥38 on PCL5 and ≥26 on NSI, and (iii) 28 subjects were grouped as combat controls with a score <38 on PCL5 and <26 on NSI, no DSM-IV-TR or DSM-V diagnosis of a psychotic disorder (e.g., schizophrenia), no mTBI within the last 5 years, and no history of a moderate-to-severe TBI. All of these three groups were matched in age, race, deployment history, and education. Comparing NSL score and PCL5 score among these groups, it can be seen that NSI scores were significantly different between the PCS+PTSD group and the PTSD and control groups combined ($p = 1.32 \times 10^{-29}$). Also the PCL5 scores were significantly different between the control group and the PTSD and PCS+PTSD groups combined ($p = 3.64 \times 10^{-44}$).

Thirty-two phenotypic variables were used including ten primary Neurocognitive measures CNS-Vital Signs® [CNS-VS] measures [75] (which is a computerized neurocognitive assessment battery), seven derived CNS-VS domain scores, eight self-report psychological health measures, and seven neurocognitive measures from a second battery, the Automated Neuropsychological Assessment Metric (ANAM 4.0) (**Table 4**). The ten primary CNS-VS measures were Symbol Digit Coding [SDC; correct responses], Stroop Test [ST] (simple and complex),

TABLE 4 | Phenotypic variables selected by GA with different clustering methods (PTSD/PCS).

Phenotypic/genetic variables		Selected variables		
		DPC	OPTICS	Hierarchical
Primary CNS-VS measures	SDC correct		✓	✓
	ST simple			
	ST complex			
	SAT correct			
	SAT RT			
	CPT correct			
	CPT RT			
	DTT percent box			
	DTT correct			
	DST			
Derived CNS-VS measures	NCI			
	RT			
	VM			
	CA			
	CF			
	EF			
	PS			
Self-report measures	LC			
	PSS			
	PSQI			
	ESS			
	ZDS	✓	✓	✓
	ZAS	✓		
	CES	✓	✓	✓
	LEC		✓	✓
Other psychometric measures	CDD_SS			✓
	CDS_SS			
	MTS			
	MP			
	PRT			
	SRT2			
	SRT			

Shifting Attention Test (SAT), Continuous Performance Test [CPT; correct responses and reaction time, RT], Dual-Task Test [DTT; correct responses and RT], and Digit Span Test [DST]. Seven derived CNS-VS domain scores were verbal memory [VM], complex attention [CA], reaction time [RT], processing speed [PS], cognitive flexibility [CF], executive functioning [EF], and neurocognitive composite index [NCI], which was computed by averaging the other six domain scores. Domain scores were standardized to have a mean of 100 and standard deviation of 15. In addition, data from the ANAM seven subtests were included—Coded Digit Substitution [CDS], Coded Digit Substitution-Delayed [CDD], Matching to Sample

[MTS], Mathematical Processing [MP], Procedural Reaction Time [PRT], Simple Reaction Time [SRT], and Simple Reaction Time-Delayed [SRT2]. Effort was also assessed to improve the validity of our assessment data. Finally, the Test of Memory Malingering (TOMM) was applied consisting of two learning trials and a retention trial that uses pictures of common, everyday objects (e.g., chair, pencil). A cut-off score (<45 correct) for the first two learning trials was used to determine eligibility for participation in the study.

Psychological health was assessed using five self-report measures—Perceived Stress Scale [PPS], Pittsburgh Sleep Quality Index [PSQI], Epworth Sleepiness Scale [ESS], Zung Anxiety Scale [ZAS] and Zung Depression Scale [ZDS]; and three exposure/injury descriptive measures—Combat Exposure Scale [CES], lifetime concussions [LC], and Life Events Checklist [LEC].

The study protocol and procedures were approved by the Auburn University Institutional Review Board (IRB) and the Headquarters U.S. Army Medical Research and Material Command, IRB (HQ USAMRMC IRB).

Data Acquisition and Preprocessing

For each neuropsychiatric disorder, an rs-fMRI dataset was obtained using different scanners with different parameters. A standard preprocessing was then performed on each dataset, individually. The details of data acquisition and preprocessing are described in **Supplementary Material**: Data Acquisition and Preprocessing.

Connectivity Measures

Given the high dimensionality of whole-brain data, each rs-fMRI image was partitioned into 200 (for ADHD, AD, and ASD) or 125 (for PTSD/PCS) functionally homogenous regions of interests (ROIs) using spatially constrained spectral clustering [cc200 template] [76]. Even though the same parcellation was used on all the datasets, we ended up with only 125 regions for the PTSD/PCS dataset since we had limited brain coverage (cerebellum was excluded). The mean time series for each ROI was subsequently extracted. Deconvolution of ROI time series was then performed using the method proposed by Wu et al. [77] to obtain hidden neuronal time series [78–81]. Deconvolution was performed because fMRI is an indirect measure of neural activity that can be influenced, at least in-part, by non-neural factors which control the shape of the hemodynamic response function (HRF), and deconvolution minimizes the inter-subject and spatial variability of the HRF that could potentially give rise to false connectivity estimates [82–91].

Next, four connectivity matrices—statistic functional connectivity (SFC), variance of dynamic functional connectivity (vDFC) [92], statistic effective connectivity (SEC) [93–96], and variance of dynamic effective connectivity (vDEC) [97–99]—were computed using the latent neuronal time series.

Functional connectivity (FC) refers to the functional co-activation between two different brain regions. In this study, static functional connectivity (SFC) was evaluated using Pearson's correlation coefficient, which gives a constant measure of connection strength between two time series. Although most of studies investigate SFC assuming the connectivity is temporally stable, it has been shown that dynamic changes in FC are relevant to neuropathology [100] as well as behavioral performance in different cognitive domains in healthy individuals [92]. Hutchison and his colleagues also provided a comprehensive overview of dynamic functional connectivity (DFC) in rs-fMRI [101]. In this study, similar to our previous study [92], DFC was evaluated using a sliding windowed Pearson's correlation with variable window length. The window length was determined adaptively by timeseries stationarity assessed through the augmented Dickey-Fuller test (ADF test), which searches for the optimal window length within a specified range using stationarity of the signal as the criteria for optimization. According to Jia's [92] study, we used a liberal range of 20–140 data points in this work.

While FC is a non-directional quantity, another approach to brain connectivity modeling is effective connectivity (EC), which characterizes directional causal interactions in the brain. It gives characteristically different information from FC, i.e., the former characterizes causal influences while the latter captures co-activation, both of which have been acknowledged as distinct modes of communication in the brain [93]. We evaluated SEC using Granger causality [102–104], which quantifies the directional influence of one region over the other. We also evaluated its time-varying version, DEC [105–108], using time-varying Granger causality evaluated in a dynamic Kalman filter framework [109–112].

SFC, DFC, SEC, and DEC values were obtained between all pairs of brain regions. Variance of DFC (and DEC) were computed to obtain vDFC (and vDEC). This provides a single measure of variability of connectivity over time for every connection [84, 113]. In effect, we employed the measures of strength and temporal variability of co-activation and causality in this work. Significant group differences were obtained with each of these measures for each of the datasets using one-way ANOVAs, and only the top significant features ($p < 0.01$) were used in further clustering analysis. This was done in order to minimize the effect of noisy measurements and outliers on clustering analysis.

Clustering and Feature Selection

In order to test whether clinical diagnostic grouping was grounded in the underlying neurobiological and phenotypic clusters, the three clustering methods, i.e., hierarchical clustering [45, 114, 115], Ordering Points to Identify the Clustering Structure—OPTICS [46], and Density Peak Clustering—DPC [47], were applied on three types of features: (i) connectivity-based features: SFC, SEC, vDFC, and vDEC, (ii) clinical diagnostic measures, and (iii) phenotypic and genetic (when available) variables. In each clustering method, there are several user specified input parameters and the clustering results greatly depend on these parameters. To determine the optimal value for each input parameter, the Calinski-Harabasz (CH) index [116, 117] was applied in this work. Detailed description of each of the clustering methods as well as parameter optimization are described in section 3 of the **Supplementary Material**.

The clustering accuracy is often lower in high dimensional feature space, because most of the features may be irrelevant, redundant, or sometimes may even misguide results. Moreover, a large number of features make the clustering results difficult to interpret. Therefore, it is necessary to select a minimum subset of relevant features to achieve a meaningful cluster separation. For supervised learning, feature selection can be trivial, i.e., only the features that are related to the given cluster labels are maintained. Nevertheless, for unsupervised learning, the cluster labels are unknown. Thus, finding the relevant subset of features and clustering the subset of the data must be accomplished simultaneously. In this work, three different feature selection methods were performed, i.e., in house SFR, SFS [50], and GA [118, 119], to find the optimal subset of features. The performances of these three methods were compared in the Results section. Detailed description of these feature selection methods can be found in section 4 of the **Supplementary Material**.

Site-Specific Analysis

Since we are more interested in the similarity of the clusters obtained from different types of features such as connectivity and phenotypic variables and not *per se* in the clustering accuracy, we did not perform cross-validation during clustering. However, in order to determine the robustness of clustering and associated features, we performed site specific analysis. As discussed in **Supplementary Material**: Data Acquisition, ADHD and ASD datasets were obtained from different sites using different scanners, which might introduce inter-site variance and affect the clustering accuracy. To eliminate this variance, site-specific feature selection and clustering were individually applied on data acquired at each site. Let $S_1 = \{F_1, F_7, \cdots, F_m\}$ and $S_2 = \{F_3, F_5, \cdots, F_n\}$ represent connectivity features selected by the proposed feature selection and clustering framework from site 1 and site 2, respectively. The intersection between S_1 and S_2 was then used as the new "selected features" for the whole dataset.

Elimination of Outlier Subjects

Real-world data always suffers from different sources of noise, which can introduce outliers in the feature space. The accuracy of clustering depends vitally on the quality of the input data. Accordingly, the most feasible and direct way to improve the effectiveness of clustering is to eliminate outlier subjects from the data.

In this study, three different clustering methods (based on three distinct principles) were employed for revealing hidden structures in the data. For the same input data, different clustering methods will, in general, result in different partitions in terms of the number of clusters and the membership of clusters. It is impractical to find a single clustering method that can handle all the different types of datasets. However, it has been demonstrated that by combining results from different clustering methods into a "co-association" matrix [CM, [120, 121]], true underlying data membership can be identified with more fidelity. Inspired by this theory, we propose a new outlier subject elimination method by applying the union-find algorithm [122] on the co-association matrix so that isolated outlier subjects can

be identified, considered as noise in the dataset and eliminated from the analysis.

Given M different partitions for a given dataset with N subjects. The $N \times N$ co-association (CM) matrix is then defined as:

$$CM = \begin{bmatrix} CM_{11} & \dots & CM_{1j} & \dots & CM_{1N} \\ & & \vdots & & \\ CM_{i1} & \dots & CM_{ij} & \dots & CM_{iN} \\ & & \vdots & & \\ CM_{N1} & \dots & CM_{Nj} & \dots & CM_{NN} \end{bmatrix} \quad (1)$$

Each element in the CM matrix is computed by:

$$CM_{ij} = \frac{m_{ij}}{M} \quad (2)$$

Where CM_{ij} is the number of times subjects i and j are assigned to the same cluster among the M partitions.

With the CM matrix we define subjects i and j as a connected pair with condition $CM_{ij} = 1$, which indicates that subjects i and j are always grouped together among the M partitions. Note that CM is a symmetric matrix, thus only upper triangular (or lower triangular) connected pairs need to be considered. A union-find algorithm is then applied so that connected subjects are merged together. Given N subjects and its corresponding CM matrix, the union-find algorithm is described next (**Figure 2**):

1. Initially, each subject was considered as a singleton tree with only itself in it.
2. By looking up the CM matrix, the connected pairs were identified.
3. For one connected pair a and b, where a belongs to tree A and b belongs to tree B, a "find" operation was applied to find the root of tree A and root of tree B, respectively.
4. A union operation was then applied to merge trees A and B.
5. Step 3 and 4 were repeated for all connected pairs.

The output of the union-find algorithm was a set of trees, and those trees with only one node in it were considered as outlier subjects.

Functional Interpretation of Selected Connectivity Features—Enrichment Analysis

Interpretation of large-scale neuroimaging finds, e.g., brain connectivity, is often done by associating identified regions or connections to previous studies. Such an approach is developed based on subjective visual inspection or on percent of overlap with existing maps without any statistical justification. Therefore, it has potential risk of false positive interpretations and overlooking additional findings. In this study, to avoid these shortcomings, a functional interpretation method— enrichment analysis [123]—was employed, which provides a quantitative statistical measure on the association between selected connectivity features and pre-defined functional brain networks.

FIGURE 2 | Illustration of the outlier subject elimination process (using upper triangular connected pairs), employing the co-association matrix approach. $x_1, x_2, \cdots, x_6$ are six representative subjects used in the illustration.
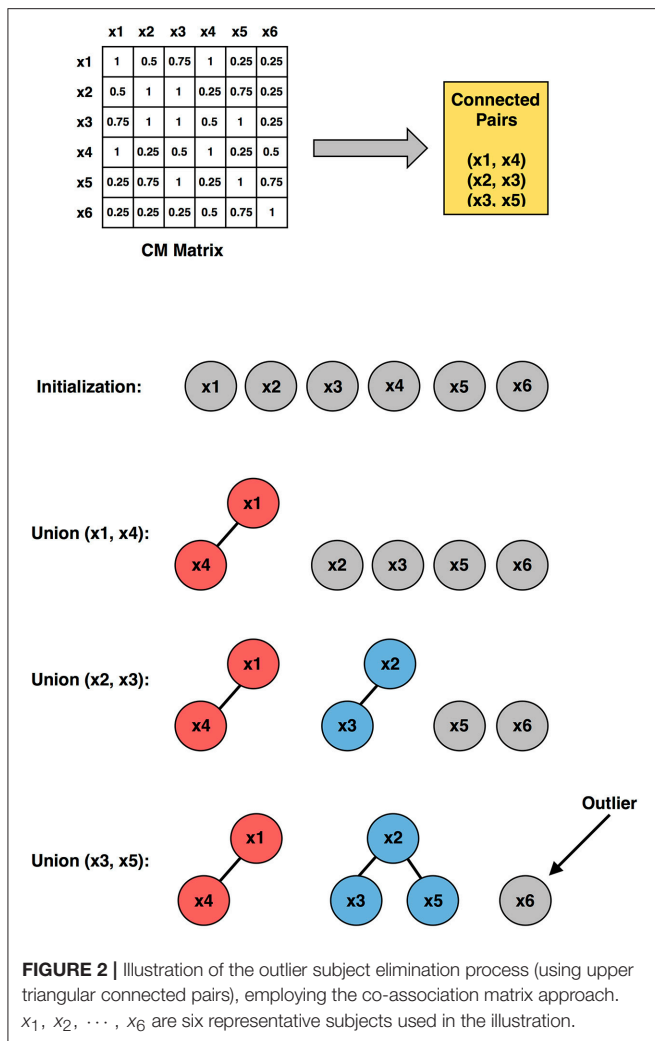
We define the following: (1) a background set S with m predefined ROIs, i.e., 200 ROIs (for ADHD, AD, and ASD) or 125 ROIs (for PTSD/PCS), and (2) a group of n selected connectivity features $A = \{(p_1, q_1), (p_2, q_2), \cdots, (p_n, q_n)\}$, where each p_i and q_i represents ROIs. Two disjoint subsets of S, C and D (with size m_c and m_d), were generated by enrichment analysis, each of which constitutes a known brain network identified in previous studies. A group B was then generated with all possible ROI pairs (i.e., connectivity features) between C and D. The size of B was determined by $K = m_c \times m_d$. Let x represent the intersection between A and B. The significance of x is the probability of having x or more elements in the intersection, which can be calculated by,

$$ p = F(x|M, n, K) = \sum_{i=x}^{\min(n, K)} \frac{\binom{K}{i}\binom{M-K}{n-i}}{\binom{M}{n}} \tag{3} $$

Where $M = \frac{m(m-1)}{2}$ is the total pairs of ROIs in the background set S. Equation **(14)** is the so called hypergeometric

(HG) cumulative distribution, which is equivalent to a one-tailed Fisher's exact test. The underlying null hypothesis of this test is that A was randomly selected from the set of all groups of ROI pairs with the same number of connectivities n over the same set of ROIs. By using this method, the statistically significant brain network-to-network (N2N) connections can be verified and quantified with corresponding p-values.

The entire pipeline for identifying different brain-based disorders, along with several supplementary analyses, e.g., site-specific analysis, elimination of outlier subjects, and enrichment analysis, is illustrated in **Figure 1**.

RESULTS

The optimal values of each input parameter determined for the three clustering methods are presented in **Tables 5**, **6**. Different feature selection methods were compared in terms of peak similarity obtained for the different neuropsychiatric disorders. From **Tables 7–10**, it can be seen that the minimum subset of features selected by GA consistently resulted in the highest similarity between clusters obtained from clinical diagnoses, fMRI-based connectivity and phenotypic variables. Using GA, the average and maximum similarities between connectivity and clinical diagnosis were 80.59 and 100%, respectively, the average and maximum similarities between connectivity and phenotypic variables were 76.72 and 80.38%, respectively, and the average and maximum similarities between clinical diagnosis and phenotypic variables were 73.06 and 76.62%, respectively. SFS was less reliable than GA in that the average and maximum similarities achieved between connectivity and clinical diagnosis were 72.20 and 100%, respectively; and the average and maximum similarities between connectivity and phenotypic variables were 66.95 and 72.22%, respectively. For similarity, the number of features determined by SFS was larger than that selected by GA. For instance, in the PTSD/PCS dataset, although the peak similarities obtained by using SFS and GA with OPTICS were similar, the number of features selected by these two methods were 84 and 15, respectively. The similarities obtained by SFR were much lower than that obtained by SFS and GA, and the number of clusters determined using SFR was different from that using SFS and GA in all datasets. The convergence of SFR, SFS, and GA were also compared. In **Figure 3**, the similarity between connectivity and phenotypic variables obtained using hierarchical clustering and different feature selection methods was plotted as a function of the number of iterations in ADHD dataset. The shape of the curve looks comparable between connectivity and clinical diagnosis for the different clustering methods, but the amplitude may be different. With GA and SFS, a clearly step-wise convergence was observed. Although SFS converged faster than GA, a lower similarity was achieved after the curve became stable. With SFR, no clear convergence was observed (i.e., the curve oscillated dramatically).

The performance of the different clustering methods varied across the datasets. Hierarchical clustering gave higher similarity in ADHD (**Table 7**) and ASD (**Table 8**) datasets. OPTICS performed better in AD (**Table 9**) and PTSD/PCS (**Table 10**)

datasets. DPC also resulted in a higher similarity in PTSD/PCS. The computation time of DPC was longer than hierarchical and OPTICS. For example, using 2.3 GHz Intel Core i7 processor, the computing time for one iteration using the PTSD dataset were as follows: hierarchical clustering took 0.27 s, OPTICS took 0.42 s, and DPC took 5.22 s, due to the fact that more input parameters (ρ and δ) were required to be optimized in DPC than that in hierarchical (cutting height) and OPTICS (threshold of reachability plot). More parameters result in larger searching space.

Site-specific analysis was applied on the ADHD dataset. We could not apply this analysis on ASD dataset since there was only one site that had enough samples for HC and disease subgroups (**Table 11**) whereas AD and PTSD datasets were obtained on the same scanner. For ADHD dataset, NYU and Peking had more than 30 samples for control, ADHD-C, and ADHD-I (**Table 12**). Thus, a site-specific analysis was applied on these two sites, individually.

The peak similarity obtained between clinical diagnostic and phenotypic clusters, between clinical diagnostic and connectivity clusters, and between phenotypic and connectivity clusters for site-specific analysis are shown in **Table 13**. Compared with previous results presented in **Table 7** using feature selection and clustering on the entire dataset across different sites, the similarity was increased by applying site-specific analysis for Peking and NYU, individually. The similarity was reduced by applying clustering on the whole datasets with commonly selected features from these two sites.

For ADHD and AD, highest similarity was achieved between connectivity and phenotypic clusters and the corresponding similarity between clinical diagnostic and phenotypic clusters was lower. On the other hand, for ASD and PTSD/PCS, highest similarity was achieved between connectivity and clinical diagnostic clusters. This suggests that diagnostic criteria for ASD and PTSD/PCS are mapped well onto underlying neurobiological clusters, while that was not the case for ADHD and AD. Consequently, for ADHD and AD, we reassigned diagnostic labels based on those generated by connectivity clusters to form new neurobiologically-informed groups. In order to verify whether this new grouping is valid, we estimated the statistical separation of phenotypic variables based on the traditional diagnostic grouping as well as with the new neurobiologically informed groups. The results, shown in **Figures 4**, **5**, indicates that almost all *p*-values were smaller with the new grouping by conducting 2-sample *t*-test. This suggests that when traditional diagnostic groups do not map well onto underlying neurobiological clusters, connectivity can be used to regroup the subjects so that they map better onto the behavioral phenotypes.

The peak similarity obtained with and without outlier subject elimination was compared and is shown in **Table 14**. Consistently higher similarity was achieved by removing the identified outlier subjects from the dataset. Moreover, in AD dataset, the number of clusters identified by hierarchical clustering was changed from

TABLE 5 | Estimated optimal values of each input parameter in clustering for clinical vs. connectivity comparison.

Disease name	DPC		OPTICS	Hierarchical
	ρ	δ	Reachability threshold	Cutting height
ADHD	46.93	1.22	1.03	1.15
AD	6.37	1.17	1.11	1.17
ASD	22.65	1.06	1.18	1.16
PTSD/PCS	12.05	1.17	0.42	1.16

TABLE 6 | Estimated optimal values of each input parameter in clustering for phenotypic vs. connectivity comparison.

Disease name	DPC		OPTICS	Hierarchical
	ρ	δ	Reachability threshold	Cutting height
ADHD	37.64	1.19	1.21	1.14
AD	6.00	1.18	1.18	1.16
ASD	26.54	1.02	1.11	1.16
PTSD/PCS	14.89	1.11	0.40	1.08

TABLE 7 | Peak similarity (highlighted), corresponding number of features, and number of clusters obtained using SFR, SFS, and GA with different clustering methods for ADHD dataset.

Feature selection method	Clustering method	Clinical vs. phenotypic			Clinical vs. connectivity			Phenotypic vs. connectivity		
		Peak Sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters
SFR	DPC	66.19	2	3	48.91	3	2	49.94	3	2
	OPTICS	64.19	2	2	56.10	4	3	58.22	4	3
	Hierarchical	61.41	3	2	51.38	195	2	51.84	18	2
SFS	DPC	66.19	2	3	58.36	37	3	68.30	17	3
	OPTICS	64.19	2	2	58.40	13	3	66.46	26	3
	Hierarchical	67.23	3	3	59.32	54	3	69.79	51	3
GA	DPC	73.51	4	3	62.48	72	3	63.83	63	3
	OPTICS	70.52	3	3	64.44	122	3	64.42	25	3
	Hierarchical	69.29	5	3	69.34	83	3	74.61	121	3

TABLE 8 | Peak similarity (highlighted), corresponding number of features, and number of clusters obtained using SFR, SFS, and GA with different clustering methods for AD dataset.

Feature selection method	Clustering method	Clinical vs. phenotypic			Clinical vs. connectivity			Phenotypic vs. connectivity		
		Peak sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters
SFR	DPC	64.46	2	4	56.68	61	4	57.99	175	3
	OPTICS	64.16	2	3	66.35	84	4	60.48	84	4
	Hierarchical	64.46	2	4	67.63	241	4	57.20	241	4
SFS	DPC	57.05	3	4	59.26	2	4	56.30	4	4
	OPTICS	65.93	4	4	64.03	20	4	57.58	12	4
	Hierarchical	65.14	4	5	64.82	155	4	60.26	193	5
GA	DPC	62.53	3	4	63.52	44	4	74.55	97	4
	OPTICS	65.93	4	4	64.23	116	4	76.75	58	4
	Hierarchical	65.14	4	5	66.65	74	4	73.28	124	5

TABLE 9 | Peak similarity (highlighted), corresponding number of features, and number of clusters obtained using SFR, SFS, and GA with different clustering methods for ASD dataset.

Feature selection method	Clustering method	Clinical vs. phenotypic			Clinical vs. connectivity			Phenotypic vs. connectivity		
		Peak sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters
SFR	DPC	66.67	9	2	58.91	5	3	61.23	5	3
	OPTICS	74.75	10	3	62.14	10	3	64.53	99	3
	Hierarchical	75.81	9	3	61.04	681	2	60.64	336	2
SFS	DPC	66.20	2	2	64.41	33	3	63.79	42	3
	OPTICS	76.00	8	3	64.50	144	3	61.75	2	3
	Hierarchical	76.00	5	3	64.79	85	3	65.52	143	3
GA	DPC	75.85	5	3	72.45	60	3	68.98	50	3
	OPTICS	76.46	5	3	79.18	103	3	74.47	70	3
	Hierarchical	76.63	5	3	89.20	101	3	75.66	54	3

TABLE 10 | Peak similarity (highlighted), corresponding number of features, and number of clusters obtained using SFR, SFS, and GA with different clustering methods for PTSD/PCS dataset.

Feature selection method	Clustering method	Clinical vs. phenotypic			Clinical vs. connectivity			Phenotypic vs. connectivity		
		Peak sim. (%)	# of features	# of clusters	Peak sim. (%)	# of features	# of clusters	Peak SIM. (%)	# of features	# of clusters
SFR	DPC	63.82	6	2	66.67	9	2	61.38	9	2
	OPTICS	69.20	4	3	66.67	16	2	76.15	18	3
	Hierarchical	61.89	6	2	66.67	6	2	57.31	4	2
SFS	DPC	74.25	3	3	73.01	2	3	70.40	4	3
	OPTICS	69.76	7	3	100	84	3	66.39	2	3
	Hierarchical	69.35	3	3	75.30	2	3	72.22	2	3
GA	DPC	74.25	3	3	100	40	3	77.25	25	3
	OPTICS	76.15	4	3	100	15	3	80.38	8	3
	Hierarchical	70.15	5	3	77.65	2	3	64.70	1	3

5 to 4 with outlier elimination (highlighted in **Table 14**), which matched with the grouping obtained using clinical diagnosis. The data in ADHD and ASD datasets comprised of data acquired at different sites using different scanners, which might explain the fact that the number of outliers identified in ADHD and ASD were generally greater than the other two datasets.

FIGURE 3 | Similarity between connectivity and phenotypic variables obtained from different iterations using hierarchical clustering and different feature selection methods: **(A)** SFR, **(B)** SFS, and **(C)** GA.
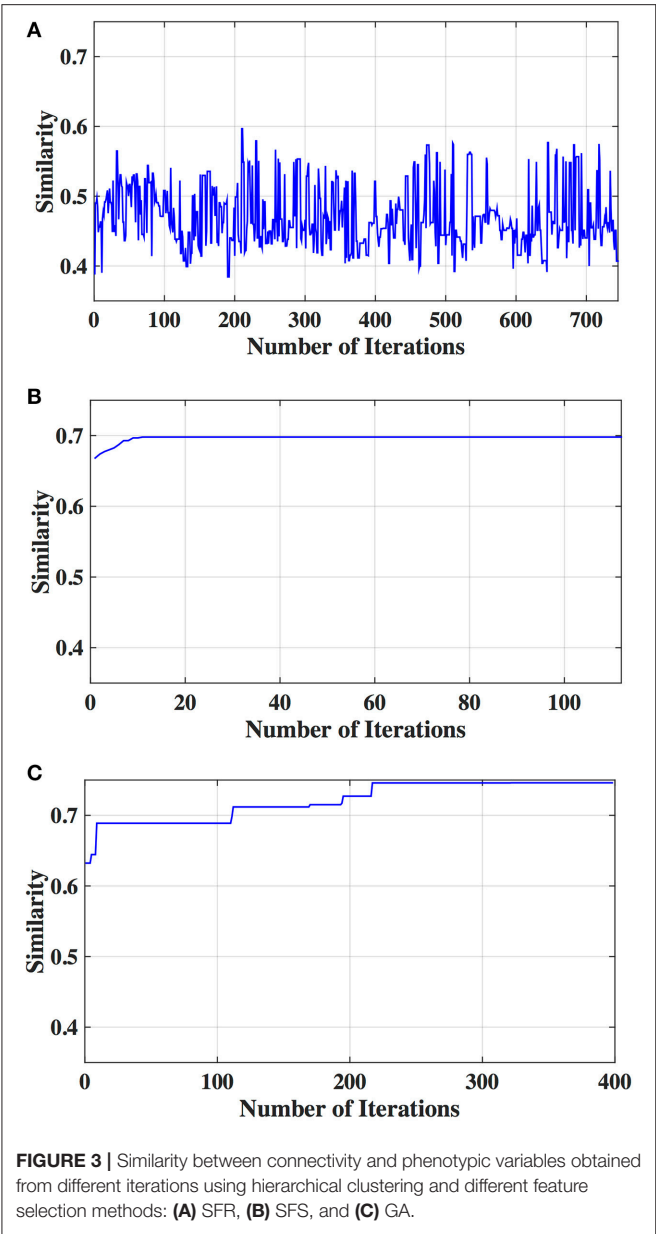
DISCUSSION

In this work, we have proposed a general analysis pipeline for characterizing different neuropsychiatric disorders using unsupervised learning methods. Our results suggest that neurobiological and phenotypic biomarkers could potentially be used as an aid by the clinician, in additional to currently available diagnostic standards, to improve diagnostic precision and identify diagnostic sub-groups. First, we discuss the selected brain connectivity features and phenotypic variables for each disorder and compare our results with previous studies. Second, we elaborate on the implications of results obtained within specific sites in comparison to those obtained from the entire ADHD dataset. Third, we discuss the reassignment of diagnostic

TABLE 11 | Number of subjects provided by each site in the ASD sample.

Site Name	HC	Autism	Asperger's
Caltech	19	13	0
CMU	13	14	0
Pitt	27	30	0
UCLA	45	44	0
SDSU	23	2	5
Trinity	24	10	6
NYU	105	53	21

TABLE 12 | Number of subjects provided by three sites in the ADHD sample.

Site name	HC	ADHD-C	ADHD-I
NYU	98	73	43
Peking	116	29	49
KKI	58	16	5

labels based on those generated by connectivity clusters. Finally, we delineate the role of outlier subject elimination in unsupervised leaning methods as applied to neuroimaging.

Connectivity Features Important for Clustering

After applying clustering, the selected connectivity features were split into two networks, i.e., (1) a network in which functional/effective connectivities and temporal variability of constituent paths were significantly ($p < 0.05$, FDR corrected) larger in the control group, and (2) a network in which functional/effective connectivities and temporal variability of the constituent paths were significantly ($p < 0.05$, FDR corrected) larger in the disease group. Here, "disease group" refers to all pathological subgroups combined. This was done since all disease groups have two or more pathological sub-groups and it becomes increasingly complex to interpret all pairwise differences. Then, these two networks were mapped back to the image space and overlaid on an anatomical glass brain (using BrainNet Viewer [124]) for the visualization, respectively. The identified brain networks were then qualitatively interpreted and compared with previous studies using enrichment analysis [123].

Intrinsic connectivity networks (ICNs) denote groups of brain regions that show correlated spontaneous activities at "resting" state [125]. It has been shown that ICNs reflect strong coupling of spontaneous fluctuations in ongoing activity and remain robust under different mental states, e.g., sleep, loss of consciousness, etc. [126, 127]. ICNs provide a common neurofunctional framework for investigating cognitive dysfunction in different neuropsychiatric disorders. There are many stable ICNs that have been identified in the human brain so far. Five of them—default mode network (DMN), visual network (VN), basal ganglia network (BGN), sensory motor network (SMN), and the semantic cognition and attention (SCAN)—have been

TABLE 13 | Similarity achieved using data from individual sites and for the whole dataset using features commonly selected by NYU and Peking.

Site name	Clinical vs. phenotypic			Clinical vs. connectivity			Phenotypic vs. connectivity		
	Hierarchical	OPTICS	DPC	Hierarchical	OPTICS	DPC	Hierarchical	OPTICS	DPC
NYU	74.32	75.87	77.02	100	100	79.38	87.02	87.02	74.55
Peking	78.99	66.67	82.18	74.66	89.02	88.83	82.32	89.84	88.42
Whole dataset	63.27	65.14	66.95	62.92	58.37	54.69	71.54	59.38	54.79
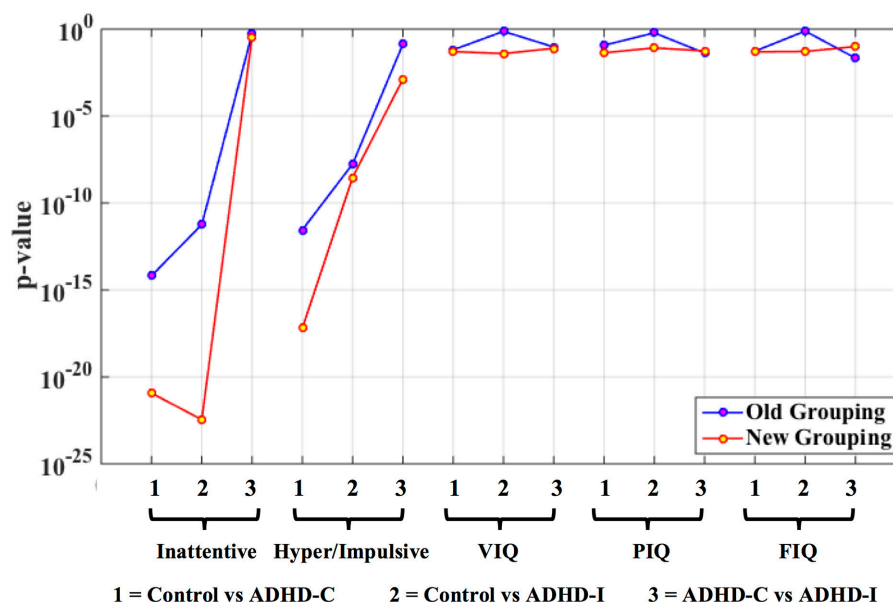
FIGURE 4 | Statistical significance (p-value) of selected phenotypic/genotypic variables with both the traditional clinical diagnostic grouping and the new connectivity grouping. The results are shown here for the ADHD dataset. Logarithmic scale is used for the y-axis of p-values.

demonstrated to be particularly important for understanding higher cognitive function and dysfunction, and provide useful models for identifying rs-fMRI connectivity patterns. Below, we discuss the significance of each of these networks to provide a context for presenting alterations in the interactions within and between these networks observed in neuropsychiatric disorders.

The DMN is one of the most well-known ICNs, which is a distributed network anchored in the posterior cingulate cortex [PCC], medial prefrontal cortex [mPFC], medial temporal lobe [MTL], precuneus, anterior cingulate cortex [ACC], inferior parietal lobe [IPL], and medial orbital gyrus [MOG] [128]. The PCC, hippocampus, and angular gyrus are typically associated with episodic memory retrieval [129, 130], autobiographical memory [131], and semantic memory related to internal thought [132]. mPFC has been demonstrated to be associated with self-related and social cognitive processes [133], value-based decision making [134], and emotion regulation [135]. Together, the entire DMN comprises an integrated system involving episodic memory, autobiographical memory, and self-related mental processes.

The VN [136] involves the occipital and bilateral temporal regions including the middle occipital gyrus, inferior temporal gyrus [ITG], fusiform gyrus, and cuneus, which is involved in visual processing and mental imagery [137, 138]. The middle occipital gyrus, ITG, and fusiform gyrus are primarily involved in the higher functions of vision processing, e.g., distinguishing objects among different categories, face recognition, visual words recognition, representation of complex object features, etc. [139, 140]. The cuneus has been demonstrated to be involved in basic visual processing, which receives visual information from retina [141].

The BGN is predominantly located in the basal ganglia including the striatum (which is subdivided into the caudate nucleus and putamen), globus pallidus or pallidum substantia nigra and thalamus [142]. The BGN is associated with a variety of functions including control of voluntary motor movements [143], procedural learning, eye movements [144], cognition [145], emotion [146], etc.

The SMN involves the precentral gyrus, postcentral gyrus, cerebellum, posterior insula, and part of the frontal gyrus corresponding to the primary sensory motor cortex and supplementary motor area [SMA] [147, 148]. Studies have indicated that this network is processing somatosensory stimuli, executing motor movements and sensorimotor integration [149, 150].
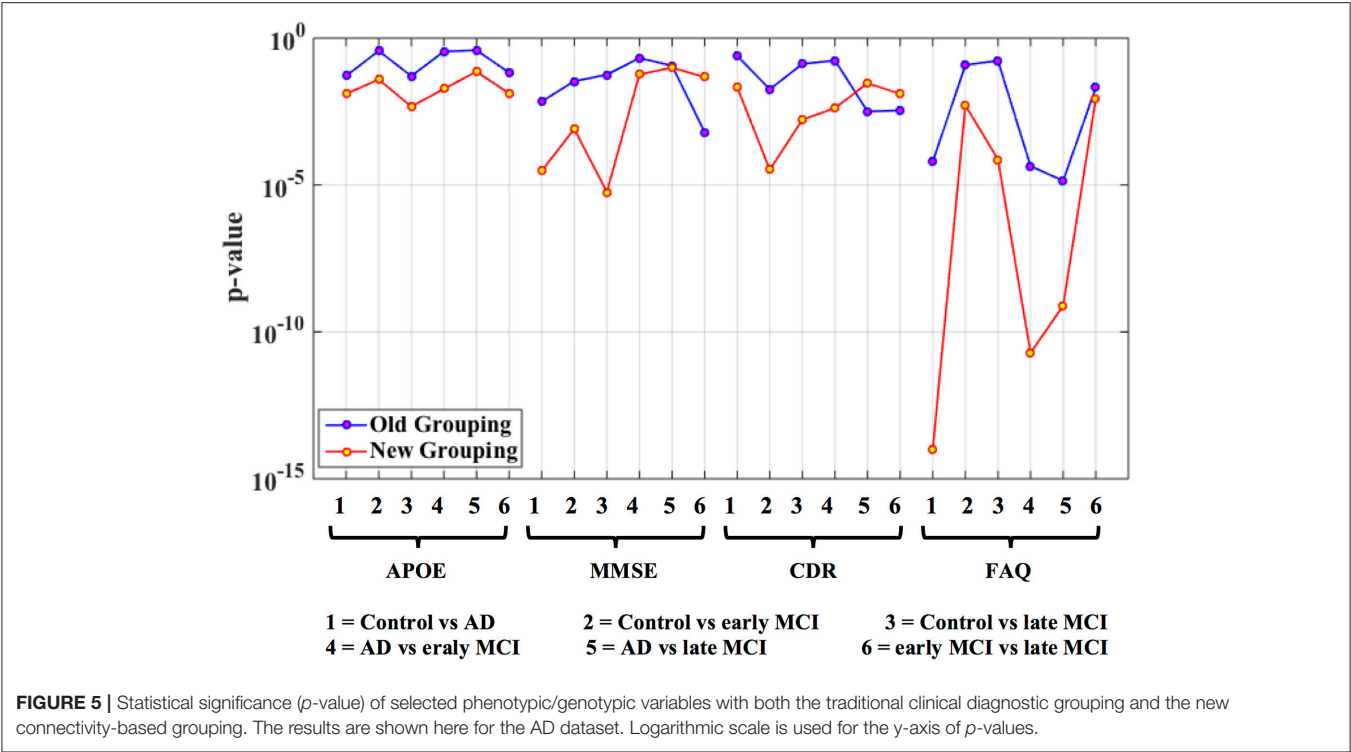
FIGURE 5 | Statistical significance (*p*-value) of selected phenotypic/genotypic variables with both the traditional clinical diagnostic grouping and the new connectivity-based grouping. The results are shown here for the AD dataset. Logarithmic scale is used for the y-axis of *p*-values.

TABLE 14 | Comparison of peak similarity obtained with and without elimination of outlier subjects.

Disease name	Clustering method	With outlier subjects removed								Including all subjects			
		Clinical vs. connectivity				Phenotypic vs. connectivity				Clinical vs. connectivity		Phenotypic vs. connectivity	
		p	*p* (%)	sim	*k*	*P*	*p* (%)	sim	*k*	sim	*k*	sim	*k*
ADHD	DPC	54	11.09	69.72	3	38	7.80	64.86	3	62.48	3	63.83	3
	OPTICS			69.79	3			68.92	3	64.44	3	64.42	3
	Hierarchical			69.62	3			75.66	3	69.34	3	74.61	3
AD	DPC	14	14.58	65.33	4	9	9.38	76.70	4	63.52	4	74.55	4
	OPTICS			65.88	4			77.75	4	64.23	4	76.75	4
	Hierarchical			69.68	4			78.38	4	66.65	5	73.28	5
ASD	DPC	15	3.30	72.65	3	49	10.79	71.95	3	72.45	3	68.98	3
	OPTICS			79.83	3			73.63	3	79.18	3	74.47	3
	Hierarchical			89.84	3			76.05	3	89.20	3	75.66	3
PTSD	DPC	3	3.45	100	3	5	5.57	77.57	3	100	3	77.25	3
	OPTICS			100	3			81.54	3	100	3	80.38	3
	Hierarchical			80.60	3			72.54	3	77.65	3	64.70	3

p, number of outliers; k, number of clusters; sim, clustering similarity. In AD dataset, the number of clusters identified by hierarchical clustering was changed from 5 to 4 with outlier elimination (highlighted), which matched with the grouping obtained using clinical diagnosis.

The SCAN is defined as regions associated with the semantic cognition network and attention network, which is a network of lateral structures in the frontal and parietal cortices, as well as some temporal regions. The semantic cognition network is primarily made up of three regions, Broca's area, Wernicke's area, as well as parts of the middle temporal gyrus [MTG] [151, 152]. Broca's area is generally defined as comprising Brodmann areas 44 and 45. Area 44 (the posterior part of the inferior frontal gyrus [IFG]) is involved in phonological processing and language production whereas area 45 (the anterior part of the IFG) engages in the semantic aspects of language. Together, Broca's area plays an important role in processing of verbal information [153]. Wernicke's area is traditionally thought to be located in the posterior part of the superior temporal gyrus [STG], which is involved in the comprehension or understanding of written and spoken language [154]. Some studies have showed that the

MTG is involved in the retrieval of lexical syntactic information [155]. The attention network is commonly segregated into two distinct networks: a bilateral dorsal attention network (DAN), which includes the dorsal frontal and parietal cortices, and the ventral attention network (VAN), largely right-lateralized, which includes the ventral frontal and parietal cortices [129, 156]. The DAN has been associated with goal-directed, top-down attention processes in inhibitory control, working memory and response selection, whereas the VAN is related with salience processing and mediates stimulus-driven, bottom-up attention processes [157]. Moreover, it is relevant to note that dorsal and ventral systems appear to interact not only during cognitive tasks [158, 159] but also during spontaneous activity [160]. Previous literature suggests that semantic cognition and attention are intimately related. This is also borne out by the fact that many disorders such as ADHD and ASD have simultaneous deficits in semantic cognition and attention. Therefore, we considered this as one network.

A qualitative as well as quantitative interpretation of alterations of these INCs and other related brain regions in different neuropsychiatric disorders are discussed below. For each pathology, we chose the features that gave us highest similarity between clusters obtained from clinical labels, connectivity features and phenotypic features. For ADHD and AD, highest similarity was obtained between connectivity and phenotypic clusters while for ASD and PTSD/PCS, highest similarity was obtained between clinical labels and connectivity clusters. Therefore, the features obtained in these two different scenarios have different implications. For ADHD and AD data sets, it suggests that traditional clinical diagnostic grouping may not neatly map onto neurobiological and neurobehavioral clusters. This may be because of uncertainty in clearly identifying differences between disease sub-groups in ADHD (ADHD-C and ADHD-I) and AD (EMCI, LMCI, and AD). Contrarily, for ASD and PTSD/PCS data sets, it suggests that traditional clinical diagnostic grouping may in fact map well onto at least neurobiological clusters. These facts are borne out by computing the purity of clusters obtained from connectivity features for disease sub-groups within each data set. To measure cluster purity, the clusters obtained using connectivity features were regrouped using the diagnostic label, and each subject was assigned to majority class in the current cluster. Then the accuracy was measured by counting the number of correctly assigned subjects within each cluster and took the average. The cluster purity for ADHD, AD, ASD, and PTSD/PCS were 0.73, 0.75, 0.94, and 1.00, respectively. It can be seen that ASD and PTSD/PCS data sets had high purity while for ADHD and AD, the purity of clusters for disease subgroups was qualitatively lower.

ADHD

One hundred and twenty-one relevant connectivity features were selected by GA and hierarchical clustering (since this combination gave highest similarity between connectivity and phenotypic features), which were 26 SFC, 14 vDFC, 53 SEC, and 28 vDEC. These features include connections in all lobes of the brain (**Figure 6**). With enrichment analysis, two N2N

FIGURE 6 | (A) SFC; **(B)** vDFC; **(C)** SEC; **(D)** vDEC features. selected by GA and hierarchical (ADHD). Selected features were split into two groups, i.e., (1) control > disease (ADHD-C and ADHD-I) and (2) disease > control. DMN, Default mode network; VN, Visual network; BGN, Basal ganglia network; SMN, Sensory motor network; SCAN, Semantic cognition and attention network.
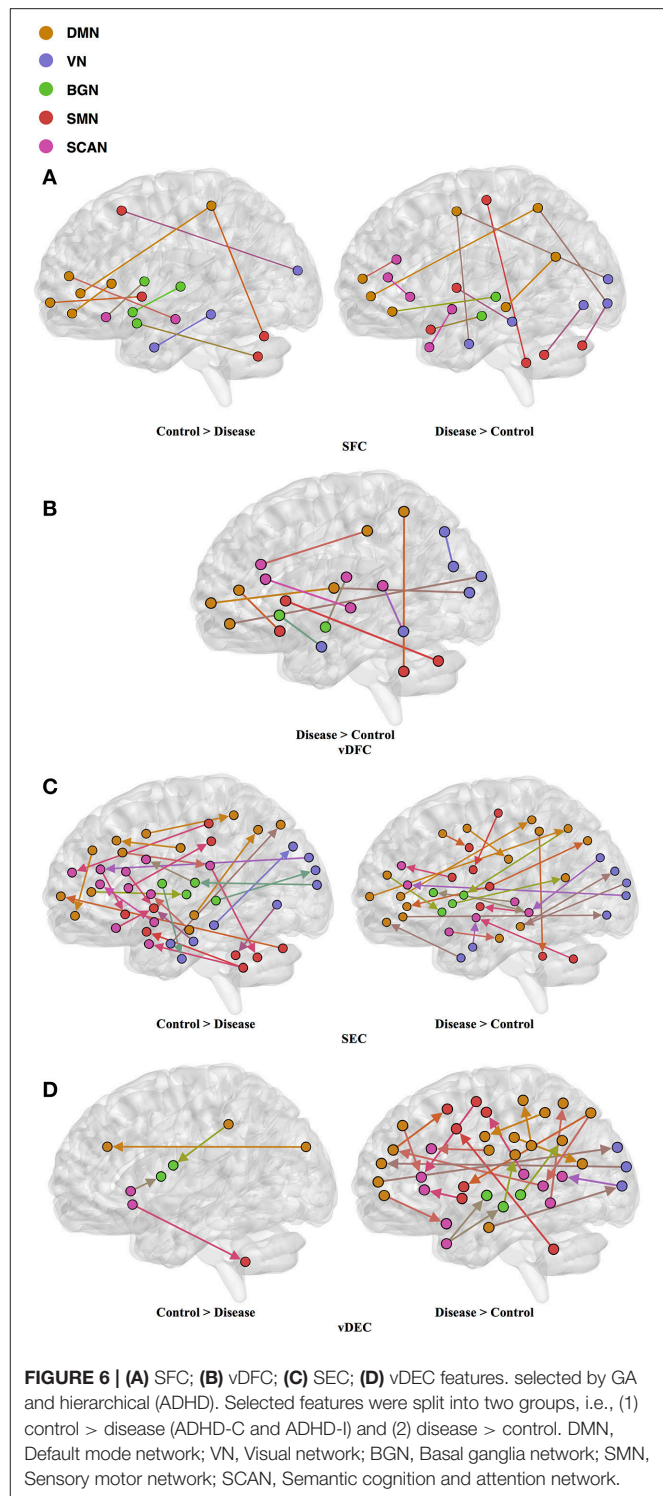
interactions were selected for SFC, i.e., the interactions within the BGN and the interaction between the VN and SMN, including connections between the cerebellum and occipital lobe, between the insula and fusiform, and between the caudate and thalamus. In addition, two N2N interactions were selected for the SEC, i.e., from the BGN to VN, and from the SCAN to SMN, including

TABLE 15 | Network-to-network interactions selected by enrichment analysis for ADHD dataset.

Feature type	Comparison	*P*-value	Selected features
SFC	Control > Disease	0.006	BGN—BGN
SFC	Disease > Control	0.04	VN—SMN
SEC	Control > Disease	0.03	BGNVN
SEC	Control > Disease	0.007	SCANSMN

connections from the caudate to occipital lobe and ITG, from the IFG and MFG to posterior insula, from the IFG to postcentral gyrus, and from the STG to cerebellum (**Table 15**).

Most of the rs-fMRI studies have demonstrated atypical functional activations in the frontal, temporal, parietal lobes, and cerebellar regions [161–163] in ADHD. Multiple studies have found aberrant functional connectivity among the brain regions of the DMN, SCAN, and BGN [164–167]. Abnormal functional activations in the orbitofrontal cortex [OFC] have been suggested to influence behavioral inhibition in children with ADHD [168]. Resting-state fMRI studies have frequently reported disrupted functional connectivity between the ACC and PCC in ADHD [169, 170]. Significantly decreased activations have been reported in the PFC, SPL, and IFG in ADHD, during multiple cognitive performance tasks and in resting-state [163, 171, 172]. One fMRI study conducted in adults with childhood ADHD showed reduced activations in bilateral IFG, left parietal lobe, caudate, and thalamus [162]. Another study found reduced functional connectivity between thalamus and other BGN areas (e.g., putamen, caudate) with ADHD [167]. Some studies have also identified reduced activations in the IFG [173] and STG [174] in ADHD patients. Kessler et al. [175] observed reduced connectivity between the SCAN and SMN and increased connectivity within the VN by applying joint independent component analysis on the ADHD-200 sample. On the other hand, increased functional connectivity in the DMN, BGN, SMN, and VN has been observed in some studies [176, 177]. Significantly increased functional connectivity between the ACC and the thalamus, cerebellum, and insula have been shown during resting-state in children with ADHD, compared to controls [170, 178, 179]. Li et al. [166] found increased connectivity between the right pulvinar and occipital regions, during a visual sustained attention task-based fMRI study. Hale et al. [180] also observed reduced activations in the VN and DMN, during letter and location judgment tasks. The features selected by GA for maximizing the similarity between connectivity and phenotypic clusters and the subset of significant N2N interactions determined by enrichment analysis are in agreement with previous literature implicating the very same regions and connections in ADHD.

AD

Fifty-eight features were selected by GA and OPTICS (since this combination gave highest similarity between connectivity and phenotypic features), including 32 vDFC features and 26 SFC features. Most of features were related to the DMN, VN, SMN, and SCAN (**Figure 7**). With enrichment analysis, two
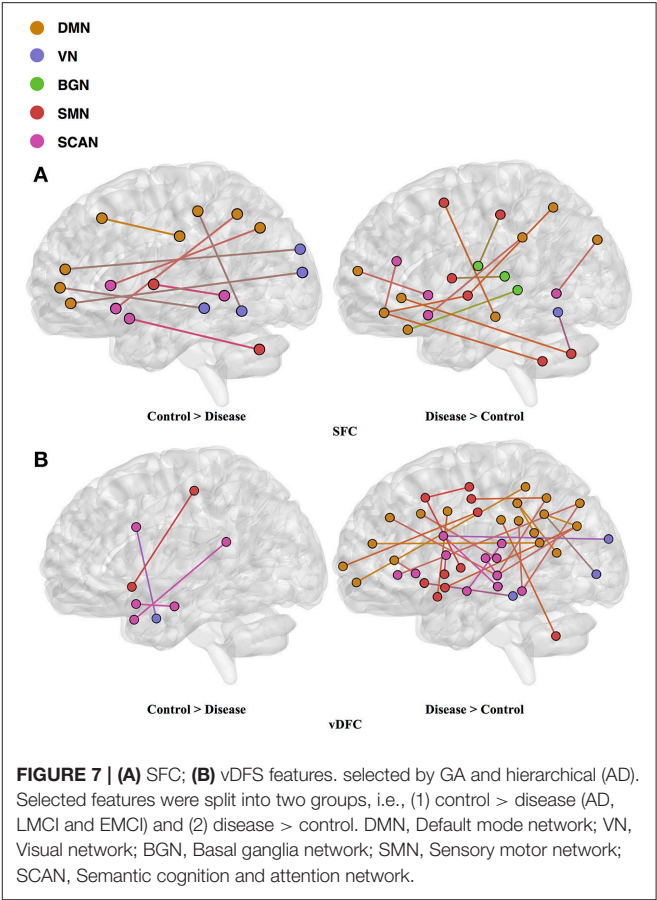

FIGURE 7 | (A) SFC; **(B)** vDFS features. selected by GA and hierarchical (AD). Selected features were split into two groups, i.e., (1) control > disease (AD, LMCI and EMCI) and (2) disease > control. DMN, Default mode network; VN, Visual network; BGN, Basal ganglia network; SMN, Sensory motor network; SCAN, Semantic cognition and attention network.

TABLE 16 | Network-to-network interactions selected by enrichment analysis for AD dataset.

Feature type	Comparison	*P*-value	Selected features
SFC	Disease > Control	0.036	DMN—SMN
SFC	Control > Disease	0.01	DMN—VN
vDFC	Control > Disease	0.01	SCAN—SCAN
vDFC	Disease > Control	0.04	DMN—SCAN

N2N interactions were selected for the SFC, i.e., the interaction between the DMN and SMN, and that between the DMN and VN, including connections between the ACC and middle occipital gyrus, between the PFC and fusiform, between the IPL and ITG, between the SFG and insula, between the hippocampus and SMA, between the cerebellum and SFG, and between the cerebellum and PFC. In addition, two N2N interactions were selected for vDFC, i.e., the interactions within the SCAN, and between the DMN and SCAN, including connections between the MTG and STG, between the PFC and IFG, between the precuneus and IFG, between the precuneus and MTG, between the PFC and STG, and between the MTG and IPL (**Table 16**).
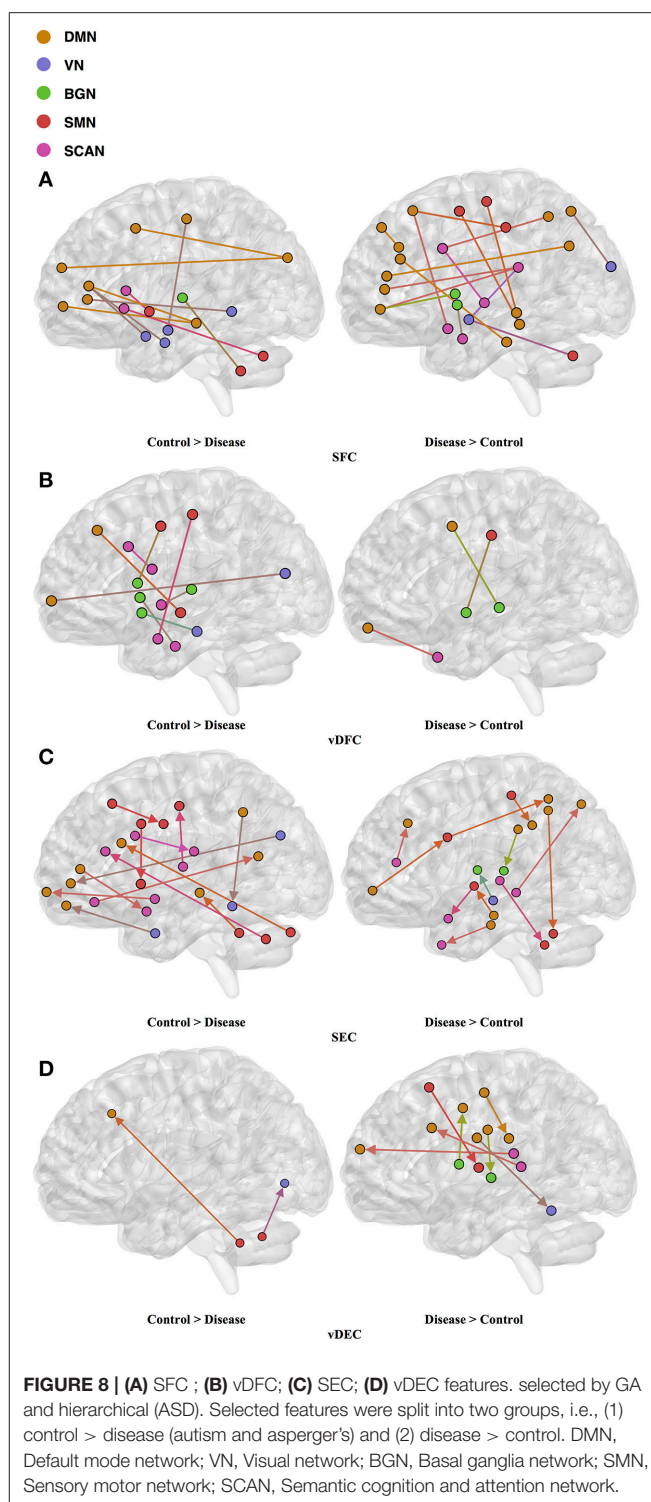
Several previous studies have indicated dysfunctions in different regions of the DMN, VN, SMN and SCAN in the AD

and MCI populations [181, 182]. Some studies have observed decreased connectivity in the DMN coupled with an increased connectivity within prefrontal regions [183–185]. Significant alterations of connectivity in the MTG, PCC, hippocampus, and angular gyrus, have been observed in AD [130, 186]. The dysfunction in the MTG, which is referred to as a central hub of the SCAN [187], is suggested as an early feature of AD [188]. A lesser degree of the MTG activation has been observed in MCI [189, 190] compared to controls. The medial parietal cortex, including the PCC and precuneus, are selectively vulnerable to amyloid deposition in AD [187], and studies of cortical metabolism using positron emission tomography and single photon emission computed tomography in AD suggest that abnormalities in the PCC and precuneus are early features of AD [188]. A voxel-based study showed that AD patients had both decreased activity of the right MFG and an increased activity of the right parietal cortex [191]. Reduced connectivity in the temporal lobe was also observed in different rs-fMRI studies [182, 184]. Multiple studies have suggested that the insula is involved in AD [192–194] and some of the behavioral abnormalities in AD may reflect insular pathology. Brier et al. [38] observed reduced anti-correlations between the DMN and SMN, and between the DMN and SCAN, during a rs-fMRI study. Li et al. [195] also found aberrant connectivity between the DMN and SCAN, as well as between the DMN and SMN. These previous studies seem to support our findings regarding features which are important for unsupervised clustering of control, EMCI, LMCI and AD groups.

ASD

Seventy-six features were selected using GA and hierarchical (since this combination gave highest similarity between connectivity features and clinical diagnosis)—30 SFC, 11 vDFC, 27 SEC, and 9 vDEC—involving the frontal, parietal, temporal lobes, and cerebellar regions (**Figure 8**). With enrichment analysis, two N2N interactions were selected for the SFC, i.e., the interaction within the DMN, and between the DMN and VN, including connections between the PFC and angular gyrus, between the SFG and angular gyrus, between the ACC and parahippocampal gyrus, between the MOG and parahippocampal gyrus, between the ACC and fusiform, between the SFG and ITG, and between the ACC and ITG. One N2N interaction was selected for the vDFC, i.e., the interaction between the BGN and SCAN, including connections between the caudate and MTG, and between the thalamus and STG. In addition, one N2N interaction was selected for the SEC, i.e., from the DMN to SMN, including connections from the MOG to precentral, from the hippocampus to posterior insula, and from the precuneus to cerebellum (**Table 17**).

Several recent studies have observed abnormal connectivity in the DMN, SCAN, SMN, BGN and VN in the pathophysiology of ASD [196, 197]. A recent meta-analysis showed alterations in the MTG, hippocampus, as well as the posterior medial cortex in ASD [198], which were suggested to be related to deficits in social information processing. It has been shown that the PCC and mPFC in ASD are hypoactive compared with healthy

FIGURE 8 | (A) SFC ; **(B)** vDFC; **(C)** SEC; **(D)** vDEC features. selected by GA and hierarchical (ASD). Selected features were split into two groups, i.e., (1) control > disease (autism and asperger's) and (2) disease > control. DMN, Default mode network; VN, Visual network; BGN, Basal ganglia network; SMN, Sensory motor network; SCAN, Semantic cognition and attention network.

controls [199]. Decreased connectivity between the PCC and SFG, the PCC and temporal lobes, as well as the PCC and parahippocampal gyri were observed, which were associated with poor social skills [200]. Dysfunction in the SCAN has been shown to be related to deficits in language and communication in individuals with ASD. Reduced activation and functional

TABLE 17 | Network-to-network interactions selected by enrichment analysis for ASD dataset.

Feature type	Comparison	*P*-value	Selected features
SFC	Control > Disease	0.033	DMN—DMN
SFC	Control > Disease	0.015	DMN—VN
vDFC	Control > Disease	0.04	BGN—SCAN
SEC	Disease > Control	0.025	DMNSMN

connectivity in the frontal-temporal SCAN were observed by Mody et al. [201]. A recent rs-fMRI study found a marked loss of functional connectivity between the right cerebellar region and regions in the SCAN [202]. Weaker connection between the SMA and ventral premotor cortex was found in the ASD group compared with controls, which has been hypothesized to underlie the initiation of speech motor actions [203]. Decreased connectivity between the BGN and the occipital region and prefrontal cortical regions was also found by Prat et al. [204]. A meta-analysis identified the posterior insula as a consistent locus of hypoactivity in ASD [199]. Other fMRI studies have also suggested that the insula is one possible key dysfunctional area in ASD [205]. In contrast, a recent rs-fMRI study [206] observed stronger functional connectivity within several large-scale brain networks in children with ASD compared with controls, including the DMN, SCAN, SMN, BGN, and VN. It has been suggested that developmental trajectories in ASD can be both heterogeneous and aberrant compared to neurotypicals and hyper- or hypo-connectivity is observed depending on when the data is acquired during development [206]. Our results are in broad agreement with previous fMRI literature in ASD discussed above.

PTSD/PCS

Fifteen features were selected by GA and OPTICS (since this combination gave the highest similarity between connectivity features and clinical diagnosis); 2 SFC, 5 vDFC, 2 SEC, and 6 vDEC. These features were mainly located in the DMN, BGN, and SCAN (**Figure 9**). With enrichment analysis, one N2N interaction between the DMN and BGN was selected for both the SFC and vDFC. This involved connections between the ACC and caudate, and between the parahippocampal gyrus and caudate. In addition, one N2N interaction from the DMN to BGN was selected for the vDEC, which included the connection from the parahippocampal gyrus to caudate (**Table 18**).

Several resting-state studies of PTSD have showed aberrant connectivity within brain structures associated with the DMN [207–209]. The parahippocampal gyri and hippocampus are critical structures in the DMN, which have been shown to be essential for memory functions, especially memorizing facts and events, and memory consolidation [210]. A previous rs-fMRI study found decreased functional connectivity in the hippocampal regions in PTSD patients [211]. The BGN has also been reported to be associated with PTSD [212–214]. PTSD has
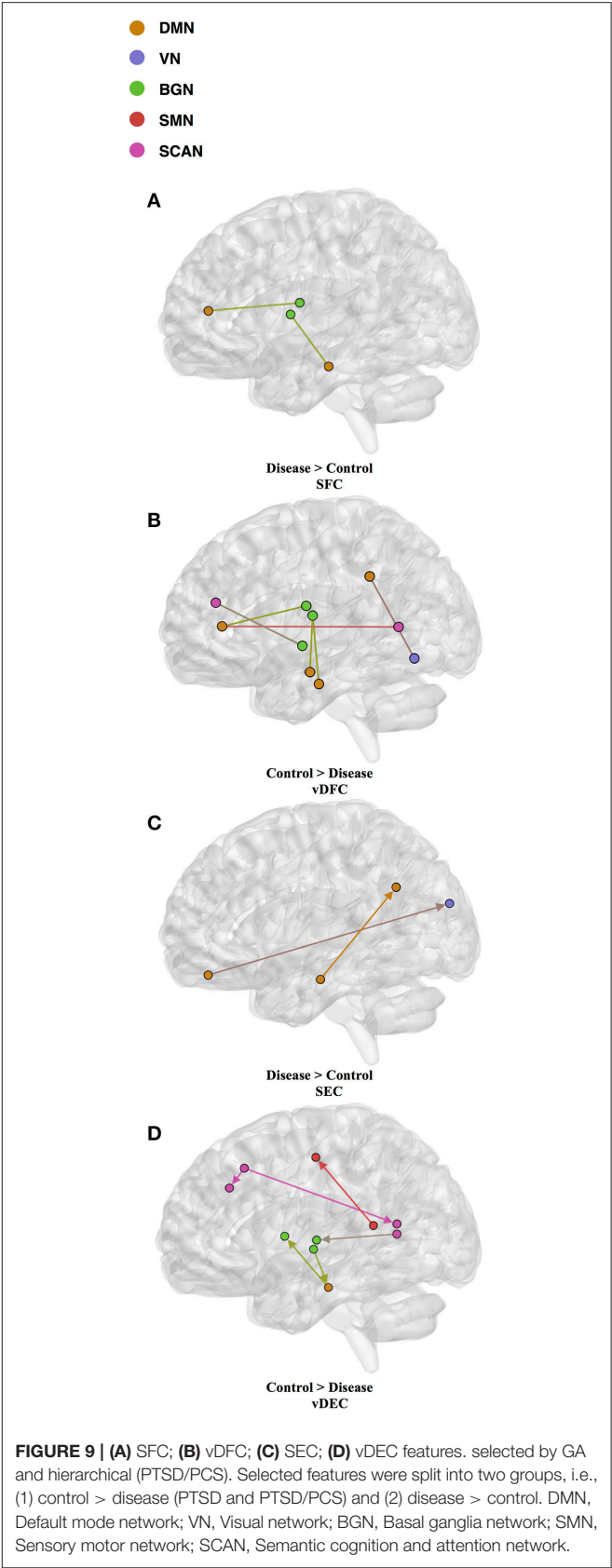

FIGURE 9 | (A) SFC; **(B)** vDFC; **(C)** SEC; **(D)** vDEC features. selected by GA and hierarchical (PTSD/PCS). Selected features were split into two groups, i.e., (1) control > disease (PTSD and PTSD/PCS) and (2) disease > control. DMN, Default mode network; VN, Visual network; BGN, Basal ganglia network; SMN, Sensory motor network; SCAN, Semantic cognition and attention network.

TABLE 18 | Network-to-network interactions selected by enrichment analysis for PTSD dataset.

Feature type	Comparison	P-value	Selected features
SFC	Disease > Control	0.001	DMN—BGN
vDFC	Control > Disease	< 0.001	DMN—BGN
vDEC	Control > Disease	0.016	DMNBGN

been linked with abnormal activation of different BGN regions, brain stem, and limbic regions compared with control groups [26, 215, 216]. The connectivity between the DMN and BGN and between the DMN and VN have been observed to be impacted in PTSD in many functional connectivity studies [217, 218]. Lanius et al. [217] found increased connectivity between the ACC and caudate, the PCC, the right parietal lobe, and the right occipital lobe, during a rs-fMRI study using subjects with PTSD. Stark et al. found changes in connectivity between the DMN and BGN, (e.g., connections between the ACC and caudate, between the parahippocampal gyri and caudate), by applying a systematic, quantitative meta-data analysis on colleagues previous studies. The SCAN has also been demonstrated to be linked to PTSD. Reduced connectivity was observed in the MTG, MFG and several BGN regions in the PTSD group, compared with controls [219]. Yin et al. [220] also found reduced connectivity in the MTG and lingual gyrus, during a rs-fMRI study. It is interesting to note that increased static connectivity and reduced variability of dynamic connectivity between the hippocampal formation and BGN regions such as the caudate has been recently reported in PTSD and PCS [84, 87] and our results seem to confirm these findings and show that those aberrations are important for unsupervised clustering of subjects into these groups.

From above discussions, it can be seen that for each individual neuropsychiatric disorder, connectivity features selected by GA with optimal clustering method are consistent with previous studies, which suggest the effectiveness of our general pipeline for identifying different brain-based disorders using unsupervised learning.

Phenotypic Features Important for Clustering

The phenotypic variables important for clustering were selected for each psychiatric disease. Below, we discuss the relevance of these variables in the context of existing literature on those measures.

ADHD

Four phenotypic variables were selected by GA and DPC including ADHD index score, Inattentive score, Hyper/Impulsive score (all are subscales in ADHD-RS), and FIQ in intelligence scale. ADHD-RS has been considered as an effective clinical diagnostic tool for assessing the severity of ADHD in children and adolescents [221, 222]. It gathers information on the severity and frequency of symptoms, the establishment of childhood onset of symptoms, the chronicity and pervasiveness of symptoms, and the impact of symptoms on major life activities. Intelligence scale has been demonstrated

to be helpful in predicting symptomatology and outcome in children with ADHD [223]. A meta-analysis showed that FIQ was lower in adults with ADHD compared to HC [63].

AD

Three phenotypic variables, i.e., MSE, CDR, and FAQ, and one genotypic variable, i.e., APOE were selected by GA and OPTICS. APOE is considered as the major genetic risk factor for AD [69]. Although the presence of APOE does not necessarily entail the development of AD, this genetic isoform probably accelerates the rate of AD conversion and progression [224]. The MMSE is the most commonly used instrument for screening memory problems and other deficits related to cognitive aging. It has been widely used to screen for dementia [64]. CDR is a global scale developed to clinically denote the presence of AD and stage its severity [225]. Several methods have been derived based on CDR to identify AD accurately; [226]. FAQ is a standardized assessment of instrumental activities of daily living, which delineates the clinical distinction between MCI and AD [227].

ASD

Five phenotypic variables were selected including ADOS_TOTAL, ADOS_COMM and ADOS_SOCIAL (all makeup ADOS test), FIQ in intelligence scale, and ADI-R_VERBAL in ADI_R test. ADOS has been extensively used in the clinic for diagnosing ASD [228, 229]. It consists of a series of structured and semi-structured presses for an interaction of specific target behaviors associated with particular tasks and by general ratings of the quality of behaviors. Further, several studies have observed higher VIQ and FIQ in ASD compared to neurotypicals; [230]. ADI-R is a structured interview conducted with the parents of the referred individual and covers the subject's full developmental history [231]. The communication and language score, as one of the three content areas in ADI-R, is useful in assessing the presence and severity of delay or total lack of language.

PTSD/PCS

Four phenotypic variables—SDC correct, ZDS, CES, and LEC—were selected by GA and OPTICS. SDC is a test of psychomotor performance, visual-motor coordination, sustained attention, and motor and mental speed, which has been shown to be related to PTSD [232]. ZDS is a short self-administered survey to quantify the depressed status of a patient. Burriss et al. [233] showed that PTSD was associated with general learning and memory impairments, and depression was considered as a mediator of these deficits. In addition, Dretsch et al. [234] revealed that depressive symptoms in individuals with PTSD account for working memory impairments. CES was constructed to measure the subjective report of wartime stressors experienced by combatants [235]. It has been demonstrated that CES is a useful tool for identifying factors associated with PTSD [236, 237]. LEC, a measure of exposure to potentially traumatic events, was developed for assisting with screening of PTSD as well. In a clinical sample of combat veterans, a significantly correlated

relationship between LEC and PTSD symptoms was observed [238].

It can be seen from the discussion above that the phenotypic (and genotypic in case of AD) variables selected by GA for maximizing the similarity of clusters obtained from them and from connectivity features indicates that they are clinically meaningful and relevant to the behavioral deficits observed in each disorder.

Site-Specific Analysis

Modern machine learning systems often integrate data from several different sources. Usually, these sources provide data of a similar type but collected under different circumstances. For example, the ADHD dataset used in this study was collected from different sites. Although fMRI images provided by these sites had similar qualities, these images were obtained from different scanners with different scanning parameters. The accuracy of machine learning algorithms can be affected by the heterogeneity of input data. To address this issue, we performed a site-specific analysis. By applying feature selection and clustering on data obtained from each individual site, the cluster similarity was increased considerably (see **Tables 7**, **13**). However, as when we applied clustering on whole dataset with commonly selected features from individual sites, the similarity was reduced. Due to inter-site variance, it is difficult for us to translate high accuracy obtained for individual site into the whole dataset. It also affects the diagnostic precision obtained from brain connectivity measures. This calls for data acquisition standards and homogenization of data acquired from different scanners.

Connectivity-Based Reassignment of Diagnostic Labels

Many brain-based disorders are highly heterogeneous, and categorization of subgroups within many disorders is yet to be completely established. Traditionally, brain-based disorders are diagnosed by clinical interviews associated with different behavioral assessments. However, it is widely acknowledged that current clinical criteria are insufficient to clearly identify most of the brain-based disorders, separate them from healthy subjects and identify sub-groups within them. Therefore, it is necessary to develop brain imaging based models for understanding how, precisely, neural circuits generate flexible behaviors and their impairments give rise to psychiatric symptoms [10]. In this study, we used unsupervised learning algorithms to discover brain connectivity-based clusters, which were not limited to existing diagnostic criteria. Instead, it focused on separating subjects into isolated clusters with maximized inter-cluster variance and minimized intra-cluster variance. After clustering, we reassigned diagnostic labels based on those generated by connectivity clusters. Compared with clinical diagnostic groups, the neurobiologically-informed groups provided better mapping from subjects to the behavioral phenotypes. This result indicates that it might be possible to view brain-based disorders from the perspective of brain connectivity measures, establishing neuroimaging-based biomarkers for different neuropsychiatric disorders.

Outlier Subject Elimination

The overarching aim of healthcare is personalized medicine. However, basing individualized treatments on brain imaging characteristics is in the nascent stages, i.e., some subjects will deviate considerably from the normative population distribution and it becomes easier to assess population level characteristics when such subjects are eliminated from the analysis. As shown in this study, with the proposed subject outlier elimination process, the precision of clustering was improved. Note that, the inter-individual variability may be introduced not just by the variability in the underlying neuropathology, but also by non-neural sources of variance such as different scanners and/or different scanning parameters. Until a standard data acquisition process is established, outlier subject elimination will serve to homogenize the data and make better inferences at the population level.

CONCLUSION

Many neuropsychiatric disorders are conventionally diagnosed based on clinical interviews and behavioral assessments. Inherent limitations of specific measures and clinical judgment contribute to a far from perfect process. Therefore, it is necessary to establish neuroimaging-based biomarkers to improve diagnostic precision and accuracy. Rs-fMRI has been used as a promising technique for characterization and classification of different disorders. However, these approaches are besieged with methodological issues such as (i) *a priori* choice of clusters needed in k-means, (ii) a stopping criterion needed in hierarchical clustering, (iii) the large dimensionality of imaging data necessitates some type of dimensionality reduction for clustering to work properly and this step is either not carried out, or carried out by preselecting features not from the structure in the data, but by some external considerations such as previous findings in a given disorder, and (iv) the clusters obtained from imaging data are seldom compared by those obtained from clinical diagnostic criteria or behavioral phenotypes.

To address these four issues, a general pipeline was derived on identifying different brain-based disorders using unsupervised clustering methods. In addition, site-specific analysis and elimination of outlier subjects were also applied to improve clustering accuracy. Three selected clustering methods were adopted on three types of features: (1) fMRI connectivity measures, (2) clinical diagnostic labels, and (3) phenotypic variables. GA based feature selection method was also applied to improve clustering accuracy. The accuracy of the clustering and feature selection was assessed by computing the similarity of clusters between all three types of features. The effectiveness of the proposed pipeline was verified on five different disorders: ADHD, AD, ASD, PTSD, and PCS. For ADHD and AD, highest similarity was achieved between connectivity and phenotypic clusters, whereas for ASD and PTSD/PCS, highest similarity was achieved between connectivity and clinical diagnostic clusters. These results suggest that neurobiological and phenotypic biomarkers could potentially be used as an aid by the clinician, in addition to currently available subjective markers, to improve diagnostic precision.

The data and source code used in this work are presented elsewhere [239]. They can also be downloaded from GitHub repository: https://github.com/xinyuzhao/identification-of-brain-based-disorders.git.

FUTURE RECOMMENDATIONS

Here we discuss some directions in which the current work could be extended. First, we have applied the proposed pipeline to four different disorders. It may be worth evaluating its performance on various other disorders such as Schizophrenia, Depression etc. Second, deriving a consensus clustering estimate from various clustering algorithms may improve the correspondence between diagnostic, imaging and phenotypic clusters. Third, the current study used cross-sectional data and hence cannot make inferences on whether unsupervised clustering can infer mechanisms that may cause sub-clusters in disorders. However, future datasets obtained using longitudinal designs may investigate this aspect. Fourth, the sample sizes used in AD and PTSD datasets are smaller compared to those used in ASD and ADHD datasets. We believe that demonstration of our method on samples of various sizes is a strength. However, specific inferences regarding AD and PTSD would obviously require larger sample sizes in the future. Finally, we believe the performance deterioration we observed when we pooled data from different sites represents one of the great challenges in applying machine learning methods to neuroimaging data. Approaches that can model this variability such that inter-site variability is minimized are needed to realize the true potential of machine learning in clinical diagnostics.

AUTHOR CONTRIBUTIONS

GD conceived the study. TD, JK, and MD obtained funding and setup the study design for the PTSD data. XZ, DR, BY, and GD performed data analysis, with XZ taking the lead. XZ primarily wrote the manuscript and all other authors contributed toward interpretation of results and editing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2018.00025/full#supplementary-material

REFERENCES

1. Filipovych R, Resnick SM, Davatzikos C. JointMMCC: joint maximum-margin classification and clustering of imaging data. *IEEE Trans Med Imaging* (2012) 31:1124–40. doi: 10.1109/TMI.2012.2186977

2. Plitt M, Barnes KA, Martin A. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage Clin.* (2014) 7:359–66. doi: 10.1016/j.nicl.2014.12.013

3. Khazaee A, Ebrahimzadeh A, and Babajani-Feremi A. Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin Neurophysiol.* (2015) 126:2132–41. doi: 10.1016/j.clinph.2015.02.060

4. Deshpande G, Li Z, Santhanam P, Coles CD, Lynch ME, Hamann S, et al. Recursive cluster elimination based support vector machine for disease state prediction using resting state functional and effective brain connectivity. *PLoS ONE* (2010) 5:e14277. doi: 10.1371/journal.pone.0014277

5. Deshpande G, Libero LE, Sreenivasan KR, Deshpande HD, Kana RK. Identification of neural connectivity signatures of autism using machine learning. *Front Hum Neurosci.* (2013) 7:670. doi: 10.3389/fnhum.2013.00670

6. Deshpande G, Wang P, Rangaprakash D, Wilamowski B. Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE Trans Cybern.* (2015) 45:2668–79. doi: 10.1109/TCYB.2014.2379621

7. Libero LE, DeRamus TP, Lahti AC, Deshpande G, Kana RK. Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates. *Cortex* (2015) 66:46–59. doi: 10.1016/j.cortex.2015.02.008

8. Chen G, Ward BD, Xie C, Li W, Chen G, Goveas JS, et al. A clustering-based method to detect functional connectivity differences. *Neuroimage* (2012) 61:56–61. doi: 10.1016/j.neuroimage.2012.02.064

9. Guttula SV, Allam A, and Gumpeny RS. Analyzing microarray data of Alzheimer's using cluster analysis to identify the biomarker genes. *Int J Alzheimers Dis.* 2012:649456 (2012). doi: 10.1155/2012/649456

10. Wang XJ, and Krystal JH. Computational psychiatry. *Neuron* (2014) 84:638–54. doi: 10.1016/j.neuron.2014.10.018

11. Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *NeuroImage* 147, 736–745. doi: 10.1016/j.neuroimage.2016.10.045

12. Yao Z. Resting-state time-varying analysis reveals aberrant variations of functional connectivity in autism. *Front Hum Neurosci.* (2016) **10**:463. doi: 10.3389/fnhum.2016.00463

13. Ellegood J, Anagnostou E, Babineau BA, Crawley JN, Lin L, Genestine M, et al. Clustering autism: using neuroanatomical differences in 26 mouse models to gain insight into the heterogeneity. *Mol Psychiatry* (2014) **20**:118–25. doi: 10.1038/mp.2014.98

14. Hrdlicka M, Dudova I, Beranova I, Lisy J, Belsan T, Neuwirth J, et al. Subtypes of autism by cluster analysis based on structural MRI data. *Eur Child Adolesc Psychiatry* (2005) **14**:138–44. doi: 10.1007/s00787-005-0453-z

15. Sloan CD, Shen L, West JD, Wishart HA, Flashman LA, Rabin LA, et al. Genetic pathway-based hierarchical clustering analysis of older adults with cognitive complaints and amnestic mild cognitive impairment using clinical and neuroimaging phenotypes. *Am J Med Genet Part B Neuropsychiatr Genet.* (2010) **153**:1060–9. doi: 10.1002/ajmg.b.31078

16. Polanczyk GV, Willcutt EG, Salum GA, Kieling C, Rohde LA. ADHD prevalence estimates across three decades: An updated systematic review and meta-regression analysis. *Int J Epidemiol.* (2014) **43**:434–42. doi: 10.1093/ije/dyt261

17. LeFever GB, Arcona AP, and Antonuccio DO. ADHD among American Schoolchildren: Evidence of Overdiagnosis and Overuse of Medication. *Sci Rev Ment Health Pract.* (2003) **2**:49–60.

18. Reitz C, Brayne C, and Mayeux R. Epidemiology of Alzheimer disease. *Nat Rev Neurol.* (2011) **7**:137–52. doi: 10.1038/nrneurol.2011.2

19. Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. *Arch Neurol.* (1999) **56**:303–8. doi: 10.1001/archneur.56.3.303

20. Launer LJ, Ross GW, Petrovitch H, Masaki K, Foley D, White LR, et al. Midlife blood pressure and dementia: the Honolulu–Asia aging study? *Neurobiol Aging* (2000) **21**:49–55. doi: 10.1016/S0197-4580(00)00096-8

21. Mayes SD, Calhoun SL, Crites DL. Does DSM-IV Asperger's disorder exist? *J Abnorm Child Psychol.* (2001) **29**:263–71. doi: 10.1023/A:1010337916636

22. Miller JN, Ozonoff S. The external validity of Asperger disorder: lack of evidence from the domain of neuropsychology. *J Abnorm Psychol.* (2000) **109**:227–38. doi: 10.1037/0021-843X.109.2.227

23. McPartland JC, Reichow B, Volkmar FR.. Sensitivity and specificity of proposed DSM-5 diagnostic criteria for autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry* (2012) **51**:368–83. doi: 10.1016/j.jaac.2012.01.007

24. Chossegros L, Hours M, Charnay P, Bernard M, Fort E, Boisson D, et al. Predictive factors of chronic post-traumatic stress disorder 6 months after a road traffic accident. *Accid Anal Prev* (2011) **43**:471–7. doi: 10.1016/j.aap.2010.10.004

25. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders 5th Edition.* Arlington, TX: American Psychiatric Association (2013). doi: 10.1176/appi.books.9780890425596.744053

26. Dretsch MN, Wood KH, Daniel TA, Katz JS, Deshpande G, Goodman A, et al. Exploring the neurocircuitry underpinning predictability to threat in soldiers with PTSD compared to deployment exposed controls. *Open Neuroimag J.* (2016) **10**:111–24. doi: 10.2174/1874440001610010111

27. Greco JA, Liberzon I. Neuroimaging of fear-associated learning. *Neuropsychopharmacology* (2016) **41**:320–34. doi: 10.1038/npp.2015.255

28. Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., and Rauch, S. L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biol. Psychiatry* 62, 446–454. doi: 10.1016/j.biopsych.2006.10.011

29. Dretsch MN, Thiel KJ, Athy JR, Irvin CR, Sirmon-Fjordbak B, Salvatore A. Mood symptoms contribute to working memory decrement in active-duty soldiers being treated for posttraumatic stress disorder. *Brain Behav.* (2012) **2**:357–64. doi: 10.1002/brb3.53

30. Dretsch MN, Silverberg ND, Iverson GL. Multiple past concussions are associated with ongoing post-concussive symptoms but not cognitive impairment in active-duty Army soldiers. *J Neurotrauma* (2015) **6**:1–6. doi: 10.1089/neu.2014.3810

31. Pape MM, Williams K, Kodosky PN, Dretsch M. The community balance and mobility scale: a pilot study detecting impairments in military service members with comorbid mild TBI and psychological health conditions. *J Head Trauma Rehabil.* (2016) **31**:339–45. doi: 10.1097/HTR.0000000000000179

32. Dretsch M, Blieberg J, Amador K, Caban J, Kelly J, Grammer G, DeGraba T. Three scoring approaches to the Neurobehavioral Symptom Inventory for measuring clinical change in service members receiving intensive treatment for combat-related mTBI. *J Head Trauma Rehabil.* (2015) **31**:23–9. doi: 10.1097/HTR.0000000000000109

33. Dretsch MN, Williams K, Emmerich T, Crynen G, Ait-Ghezala G, Chaytow H, et al. Brain-derived neurotropic factor polymorphisms, traumatic stress, mild traumatic brain injury, and combat exposure contribute to postdeployment traumatic stress. *Brain Behav.* (2016) 6:e00392. doi: 10.1002/brb3.392

34. Bryant R. Post-traumatic stress disorder vs traumatic brain injury. *Dialogues Clin Neurosci.* (2011) **13**:251–62.

35. Brodersen KH, Deserno L, Schlagenhauf F, Lin Z, Penny WD, Buhmann JM, et al. Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage Clin.* (2014) **4**:98–111. doi: 10.1016/j.nicl.2013.11.002

36. Van Dam NT, O'Connor D, Marcelle ET, Ho EJ, Cameron Craddock R, Tobe RH, et al. Data-driven phenotypic categorization for neurobiological analyses: beyond DSM-5 labels. *Biol Psychiatry* (2016) **81**:484–94. doi: 10.1016/j.biopsych.2016.06.027

37. Chen G, Ward BD, Xie C, Li W, Wu Z, Jones JL, et al. Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging. *Radiology* (2011) **259**:213–21. doi: 10.1148/radiol.10100734

38. Brier MR, Thomas JB, Snyder AZ, Benzinger TL, Zhang D, Raichle ME, et al. Loss of intranetwork and internetwork resting state functional connections with alzheimer's disease progression. *J Neurosci.* (2012) **32**:8890–99. doi: 10.1523/JNEUROSCI.5698-11.2012

39. Liu F, Guo W, Fouche, J.-P., Wang Y, Wang W, Ding J, et al. Multivariate classification of social anxiety disorder using whole brain functional connectivity. *Brain Struct Funct.* (2015) **220**:101–15. doi: 10.1007/s00429-013-0641-4

40. Tang Y, Jiang W, Liao J, Wang W, Luo A. Identifying individuals with antisocial personality disorder using resting-state fMRI. *PLoS ONE* (2013) **8**:e60652. doi: 10.1371/journal.pone.0060652

41. Zeng L-L, Shen H, Liu L, Hu D. Unsupervised classification of major depression using functional connectivity MRI. *Hum Brain Mapp.* (2014) **35**:1630–41. doi: 10.1002/hbm.22278

42. Ashikh V, Deshpande G, Rangaprakash D, Dutt DN. Clustering of dynamic functional connectivity features obtained from functional magnetic resonance imaging data. In: 2015 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Kochi: IEEE (2015). p. 308–12.

43. Biswal BB. Resting state fMRI: a personal history. *Neuroimage* (2012) **62**:938–44. doi: 10.1016/j.neuroimage.2012.01.090

44. Buckner RL, Krienen FM, Yeo BT. Opportunities and limitations of intrinsic functional connectivity MRI. *Nat Neurosci.* (2013) **16**:832–7. doi: 10.1038/nn.3423

45. Dasgupta S, Long PM. Performance guarantees for hierarchical clustering. *J Comput Syst Sci.* (2005) **70**:555–69. doi: 10.1016/j.jcss.2004.10.006

46. Ankerst M, Breunig MM, Kriegel, HP, Sander J. Optics: ordering points to identify the clustering structure. *ACM Sigmod Rec.* (1999) **28**:49–60. doi: 10.1145/304182.304187

47. Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks. *Science* (2014) **344**:1492–6. doi: 10.1126/science.1242072

48. Allen EA, Damaraju E, Plis SM, Erhardt EB, Eichele T, Calhoun VD. Tracking whole-brain connectivity dynamics in the resting state. *Cereb Cortex* (2014) **24**:663–76. doi: 10.1093/cercor/bhs352

49. Venkataraman A, Van Dijk KRA, Buckner RL, Golland P. Exploring functional connectivity in fMRI via clustering. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei. (2009). p. 441–4.

50. Dy JG, Brodley CE. Feature selection for unsupervised learning. *J Mach Learn Res.* (2004) **5**:845–89.

51. Bradley PS, Mangasarian OL. Feature selection via mathematical programming. INFORMS. *J Comput*. (1998) **10**:209–17.

52. Sharp C, Skinner D, Serekoane M, Ross MW. A qualitative study of the cultural appropriateness of the Diagnostic Interview Schedule for Children (DISC-IV) in South Africa. *Soc. Psychiatry Psychiatr. Epidemiol.* (2011) **46**:743–51. doi: 10.1007/s00127-010-0241-z

53. Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. *J Am Acad Child Adolesc Psychiatry* (1997) **36**:980–8. doi: 10.1097/00004583-199707000-00021

54. Pappas D. ADHD rating scale-IV: checklists, norms, and clinical interpretation. *J Psychoeduc Assess*. (2006) **24**:172–8. doi: 10.1177/0734282905285792

55. McGoey KE, DuPaul GJ, Haley E, Shelton TL. Parent and teacher ratings of attention-deficit/hyperactivity disorder in preschool: The ADHD rating scale-IV preschool version. *J Psychopathol Behav Assess*. (2007) **29**:269–76. doi: 10.1007/s10862-007-9048-y

56. Wechsler D. *Wechsler Intelligence Scale for Children-Revised (WISC-R), TEA Ediciones*. New York, NY: The Psychological Corporation (1974).

57. Reich W. Diagnostic interview for children and adolescents (DICA). *J Am Acad Child Adolesc Psychiatry* (2000) **39**:59–66. doi: 10.1097/00004583-200001000-00017

58. Conners CK, Sitarenios G, Parker JDA, Epstein JN. The Revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *J Abnorm Child Psychol*. (1998) **26**:257–68. doi: 10.1023/A:1022602400621

59. Wechsler D. *The Wechsler Intelligence Scale for Children, 4th Edn*. San Antonio, TX: The Psychological Corporation (2004).

60. Smith DR. Wechsler individual achievement test. In: Andrews JJW, Saklofske DH, Janzen HL, editots. *Handbook of Psychoeducational Assessment: Ability, Achievement, and Behavior in Children. A Volume in the Educational Psychology Series*. Houston, TX: Academic Press (2001). p. 169–93.

61. Canivez GL, Konold TR, Collins JM, Wilson G. Construct validity of the wechsler abbreviated scale of intelligence and wide range intelligence test: convergent and structural validity. *Sch Psychol Q*. (2009) **24**:252–65. doi: 10.1037/a0018030

62. Malfa GL, Lassi S, Bertelli M, Pallanti S, Albertini G. Detecting attention-deficit/hyperactivity disorder (ADHD) in adults with intellectual disability. The use of Conners' Adult ADHD Rating Scales (CAARS). *Res Dev Disabil*. (2008) **29**:158–64. doi: 10.1016/j.ridd.2007.02.002

63. Bridgett DJ, Walker ME. Intellectual functioning in adults with ADHD: a meta-analytic examination of full scale IQ differences between adults with and without ADHD. *Psychol Assess*. (2006) **18**:1–14. doi: 10.1037/1040-3590.18.1.1

64. Galasko D, Klauber MR, Hofstetter CR, Salmon DP, Lasker B, Thal LJ. The Mini-Mental State Examination in the early diagnosis of Alzheimer's disease. *Arch Neurol*. (1990) **47**:49–52. doi: 10.1001/archneur.1990.00530010061020

65. Burke WJ, Houston MJ, Boust SJ, Roccaforte WH. Use of the geriatric depression scale in dementia of the alzheimer type. *J Am Geriatr Soc*. (1989) **37**:856–60.

66. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* (1984) **34**:939–44. doi: 10.1212/WNL.34.7.939

67. Cummings JL. The Neuropsychiatric Inventory: assessing psychopathology in dementia patients. *Neurology* (1997) **48**:S10–6. doi: 10.1212/WNL.48.5_Suppl_6.10S

68. Sabbagh MN, Malek-Ahmadi M, Kataria R, Belden CM, Connor DJ, Pearson C, et al. The Alzheimer's questionnaire: a proof of concept study for a new informant-based dementia assessment. *J Alzheimer's Dis* (2010) **22**:1015–21. doi: 10.3233/JAD-2010-101185

69. Kim J, Basak JM, Holtzman DM. The role of apolipoprotein E in Alzheimer's disease. *Neuron* (2009) **63**:287–303. doi: 10.1016/j.neuron.2009.06.026

70. American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders, 4th Edn*. Washington, DC: American Psychiatric Association.

71. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, et al. Autism Diagnostic Observation Schedule (ADOS). *J Autism Dev Disord*. (2000) **30**:205–23.

72. Lord C, Rutter M, Couteur A. Autism diagnostic interview-revised. *J Autism Dev Disord*. (1994) **29**:30.

73. Dickstein BD, Weathers FW, Angkaw AC, Nievergelt CM, Yurgil K, Nash WP, et al. Diagnostic utility of the posttraumatic stress disorder (PTSD) checklist for identifying full and partial PTSD in active-duty military. *Assessment* (2015) **22**:289–97. doi: 10.1177/1073191114548683

74. Cicerone KD, Kalmar K. Persistent postconcussion syndrome: the structure of subjective complaints after mild traumatic brain injury. *J Head Trauma Rehabil*. (1995) **10**:1–17. doi: 10.1097/00001199-199510030-00002

75. Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Arch Clin Neuropsychol*. (2006) **21**:623–43. doi: 10.1016/j.acn.2006.05.007

76. Craddock RC, James GA, Holtzheimer PE, Hu XP, Mayberg HS. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp*. (2012) **33**:1914–28. doi: 10.1002/hbm.21333

77. Wu, GR, Liao W, Stramaglia S, Ding, JR, Chen H, Marinazzo D. A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data. *Med Image Anal*. (2013) **17**:365–74. doi: 10.1016/j.media.2013.01.003

78. Deshpande G, Hu X. Investigating effective brain connectivity from fMRI data: past findings and current issues with reference to Granger causality analysis. *Brain Connect*. (2012) **2**:235–45. doi: 10.1089/brain.2012.0091

79. Deshpande G, Sathian K, Hu X, Buckhalt J. A rigorous approach for testing the constructionist hypotheses of brain function. *Behav Brain Sci*. (2012) **35**:148–9. doi: 10.1017/S0140525X1100149X

80. Sathian K, Deshpande G, and Stilla R. Neural changes with tactile learning reflect decision-level reweighting of perceptual readout. *J Neurosci*. (2013) **33**:5387–98. doi: 10.1523/JNEUROSCI.3482-12.2013

81. Hutcheson NL, Sreenivasan KR, Deshpande G, Reid MA, Hadley J, White DM, et al. Effective connectivity during episodic memory retrieval in schizophrenia participants before and after antipsychotic medication. *Hum Brain Mapp*. (2015) **36**:1442–57. doi: 10.1002/hbm.22714

82. Handwerker DA, Ollinger JM, D'Esposito M. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* (2004) **21**:1639–51. doi: 10.1016/j.neuroimage.2003.11.029

83. Deshpande G, Sathian K, Hu X. Effect of hemodynamic variability on Granger causality analysis of fMRI. *Neuroimage* (2010) **52**:884–96. doi: 10.1016/j.neuroimage.2009.11.060

84. Rangaprakash D, Deshpande G, Daniel TA, Goodman AM, Robinson JL, Salibi N, et al. Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild traumatic brain injury and posttraumatic stress disorder. *Hum Brain Mapp*. (2017) **38**:2843–4. doi: 10.1002/hbm.23551

85. Rangaprakash D, Dretsch M, Yan W, Katz JS, Denney TS, Deshpande G. Hemodynamic variability in soldiers with trauma: implications for functional MRI connectivity studies. *NeuroImage* (2017) **16**:409–17. doi: 10.1016/j.nicl.2017.07.016

86. Rangaprakash D, Dretsch M, Yan W, Katz JS, Denney TS, Deshpande G. Hemodynamic response function parameters obtained from resting-state functional MRI data in Soldiers with trauma. *Data Brief* (2017) **14**:558–62. doi: 10.1016/j.dib.2017.07.072

87. Rangaprakash D, Dretsch M, Venkataraman A, Katz J, Denney TS, Deshpande G. Identifying disease foci from static and dynamic effective connectivity networks: illustration in Soldiers with trauma. *Hum Brain Mapp*. (2018) **39**:264–87. doi: 10.1002/hbm.23841

88. Rangaprakash D, Wu GR, Marinazzo D, Hu X, Deshpande G. Hemodynamic response function (HRF) variability confounds resting state fMRI connectivity. *Magn Reson Med*. (2018) **80**:1697–713. doi: 10.1002/mrm.27146

89. Rangaprakash D, Wu GR, Marinazzo D, Hu X, Deshpande G. Parameterized hemodynamic response function data of healthy individuals obtained from resting-state functional MRI in a 7T MRI scanner. *Data Brief* (2018) **17**:1175–9. doi: 10.1016/j.dib.2018.01.003

90. Yan W, Rangaprakash D, Deshpande G. Aberrant hemodynamic responses in Autism: implications for resting state fMRI functional connectivity studies. *NeuroImage* (2018) 19:320–30. doi: 10.1016/j.nicl.2018.04.013

91. Yan W, Rangaprakash D, Deshpande G. Hemodynamic Response function parameters obtained from resting state BOLD fMRI data in subjects with autism spectrum disorder and matched healthy controls. *Data Brief* (2018) 14:558–62. doi: 10.1016/j.dib.2018.04.126

92. Jia H, Hu X, Deshpande G. Behavioral relevance of the dynamics of the functional brain connectome. *Brain Connect.* (2014) 4:741–59. doi: 10.1089/brain.2014.0300

93. Deshpande G, Santhanam P, Hu X. Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data. *Neuroimage* (2011) 54:1043–52. doi: 10.1016/j.neuroimage.2010.09.024

94. Lacey S, Hagtvedt H, Patrick VM, Anderson A, Stilla R, Deshpande G, et al. Art for reward's sake: visual art recruits the ventral striatum. *Neuroimage* (2011) 55:420–33. doi: 10.1016/j.neuroimage.2010.11.027

95. Krueger F, Landgraf S, Van Der Meer E, Deshpande G, Hu X. Effective connectivity of the multiplication network: a functional MRI and multivariate granger causality mapping study. *Hum Brain Mapp.* (2011) 32:1419–31. doi: 10.1002/hbm.21119

96. Preusse F, van der Meer E, Deshpande G, Krueger F, Wartenburger I. Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. *Front Hum Neurosci.* (2011) 5:22. doi: 10.3389/fnhum.2011.00022

97. Grant MM, Wood K, Sreenivasan K, Wheelock M, White D, Thomas J, et al. Influence of early life stress on intra- and extra-amygdaloid causal connectivity. *Neuropsychopharmacology* (2015) 40:1–12. doi: 10.1038/npp.2015.28

98. Hampstead BM, Khoshnoodi M, Yan W, Deshpande G, Sathian K. Patterns of effective connectivity during memory encoding and retrieval differ between patients with mild cognitive impairment and healthy older adults. *Neuroimage* (2016) 124:997–1008. doi: 10.1016/j.neuroimage.2015.10.002

99. Feng C, Deshpande G, Liu C, Gu R, Luo YJ, Krueger F. Diffusion of responsibility attenuates altruistic punishment: a functional magnetic resonance imaging effective connectivity study. *Hum Brain Mapp.* (2016) 37:663–77. doi: 10.1002/hbm.23057

100. Sakoglu Ü, Pearlson GD, Kiehl KA, Wang YM, Michael AM, Calhoun VD. A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. *Magn Reson Mater Phys Biol Med.* (2010) 23:351–66. doi: 10.1007/s10334-010-0197-8

101. Hutchison RM, Womelsdorf T, Allen EA, Bandettini PA, Calhoun VD, Corbetta M, et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* (2013) 80:360–78. doi: 10.1016/j.neuroimage.2013.05.079

102. Deshpande G, LaConte S, James GA, Peltier S, Hu X. Multivariate granger causality analysis of fMRI data. *Hum Brain Mapp.* (2009) 30:1361–73. doi: 10.1002/hbm.20606

103. Deshpande G, Hu X, Lacey S, Stilla R, Sathian K. Object familiarity modulates effective connectivity during haptic shape perception. *Neuroimage* (2010) 49:1991–2000. doi: 10.1016/j.neuroimage.2009.08.052

104. Hampstead BM, Stringer AY, Stilla RF, Deshpande G, Hu X, Moore AB, et al. Activation and effective connectivity changes following explicit-memory training for face-name pairs in patients with mild cognitive impairment: a pilot study. *Neurorehabil Neural Repair* (2011) 25:210–22. doi: 10.1177/1545968310382424

105. Sathian K, Lacey S, Stilla R, Gibson GO, Deshpande G, Hu X, et al. Dual pathways for haptic and visual perception of spatial and texture information. *Neuroimage* (2011) 57:462–75. doi: 10.1016/j.neuroimage.2011.05.001

106. Kapogiannis D, Deshpande G, Krueger F, Thornburg MP, Grafman JH. Brain networks shaping religious belief. *Brain Connect.* (2014) 4:70–9. doi: 10.1089/brain.2013.0172

107. Goodyear K, Parasuraman R, Chernyak S, Visser E, Madhavan P, Deshpande G, et al. An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Soc. Neurosci.* (2016) 12:570–81. doi: 10.1080/17470919.2016.1205131

108. Liang P, Deshpande G, Zhao S, Liu J, Hu X, Li K. Altered directional connectivity between emotion network and motor network in Parkinson's

disease with depression. *Medicine (Baltimore).* (2016). 95:e4222. doi: 10.1097/MD.0000000000004222

109. Grant MM, White D, Hadley J, Hutcheson N, Shelton R, Sreenivasan K, et al. Early life trauma and directional brain connectivity within major depression. *Hum Brain Mapp.* (2014) 35:4815–26. doi: 10.1002/hbm.22514

110. Lacey S, Stilla R, Sreenivasan K, Deshpande G, Sathian K. Spatial imagery in haptic shape perception. *Neuropsychologia* (2014) 60:144–58. doi: 10.1016/j.neuropsychologia.2014.05.008

111. Bellucci G, Chernyak S, Hoffman H, Deshpande G, Monte OD, Knutson KM, et al. Effective connectivity of brain regions underlying third party punishment: functional MRI and Granger causality evidence. *Soc Neurosci.* (2016) 12:124–134. doi: 10.1080/17470919.2016.1153518

112. Wheelock MD, Sreenivasan KR, Wood KH, Ver Hoef LW, Deshpande G, Knight DC. Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *Neuroimage* (2014) 102:904–12. doi: 10.1016/j.neuroimage.2014.08.005

113. Jin T, Kim S-G. Cortical layer-dependent dynamic blood oxygenation, cerebral blood flow and cerebral blood volume responses during visual stimulation. *Neuroimage* (2008) 43:1–9. doi: 10.1016/j.neuroimage.2008.06.029

114. Liao W, Chen H, Yang Q, Lei X. Analysis of fMRI data using improved self-organizing mapping and spatio-temporal metric hierarchical clustering. *IEEE Trans Med Imaging* (2008) 27:1472–83. doi: 10.1109/TMI.2008.923987

115. Cheng D, Kannan R, Vempala S, Wang G. A divide-and-merge methodology for clustering. *ACM Trans Database Syst.* (2006) 31:1499–525. doi: 10.1145/1189769.1189779

116. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Simul Comput.* (1974) 3:1–27. doi: 10.1080/03610917408548446

117. Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S. Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern.* (2013) 43:982–94. doi: 10.1109/TSMCB.2012.2220543

118. Yang J, Honavar V. Feature subset selection using a genetic algorithm. *Patt Recognit.* (1997) 13:380.

119. Shahamat H, Pouyan AA. Feature selection using genetic algorithm for classification of schizophrenia using fMRI data. *J Artif Intell Data Min.* (2015) 3:30–7. doi: 10.5829/idosi.JAIDM.2015.03.01.04

120. Fred ALN, Jain AK. Combining multiple clusterings using evidence accumulation. *IEEE Trans Patt Anal Mach Intell.* (2005) 27:835–50. doi: 10.1109/TPAMI.2005.113

121. Fred ALN, Jain AK. Data clustering using evidence accumulation. *Object Recognit Support User Interact Serv Robot.* (2002) 4:276–80. doi: 10.1109/ICPR.2002.1047450

122. Tarjan RE. Efficiency of a good but not linear set union algorithm. *J ACM* (1975) 22:215–25. doi: 10.1145/321879.321884

123. Maron-Katz A, Amar D, Simon EB, Hendler T, Shamir R. RichMind: a tool for improved inference from large-scale neuroimaging results. *PLoS ONE* (2016) 11:e0159643. doi: 10.1371/journal.pone.0159643

124. Xia M, Wang J, He Y. BrainNet viewer: a network visualization tool for human brain connectomics. *PLoS ONE* (2013) 8:e68910. doi: 10.1371/journal.pone.0068910

125. Whitfield-Gabrieli S, Ford JM. Default mode network activity and connectivity in psychopathology. *Annu Rev Clin Psychol.* (2012) 8:49–76. doi: 10.1146/annurev-clinpsy-032511-143049

126. Menon V. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn Sci.* (2011) 15:483–506. doi: 10.1016/j.tics.2011.08.003

127. Vergun S, Deshpande AS, Meier TB, Song J, Tudorascu DL, Nair VA, et al. Characterizing functional connectivity differences in aging adults using machine learning on resting state fMRI data. *Front Comput Neurosci.* (2013) 7:38. doi: 10.3389/fncom.2013.00038

128. Buckner RL, Andrews-Hanna JR, Schacter DL. The brain's default network. *Ann NY Acad Sci.* (2008) 1124:1–38. doi: 10.1196/annals.1440.011

129. Sestieri C, Corbetta M, Romani GL, Shulman GL. Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. *J Neurosci.* (2011) 31:4407–20. doi: 10.1523/JNEUROSCI.3335-10.2011

130. Vannini P, O'Brien J, O'Keefe K, Pihlajamäki M, Laviolette P, Sperling RA. What goes down must come up: role of the posteromedial

cortices in encoding and retrieval. *Cereb Cortex* (2011) **21**:22–34. doi: 10.1093/cercor/bhq051

131. Spreng RN, Mar R, Kim ASN. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *J Cogn Neurosci.* (2009) **21**:489–510. doi: 10.1162/jocn.2008.21029

132. Binder JR, Desai RH, Graves WW, Conant LL. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb Cortex* (2009) **19**:2767–96. doi: 10.1093/cercor/bhp055

133. Amodio DM, Frith CD. Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci.* (2006) **7**:268–77. doi: 10.1038/nrn1884

134. Rangel A, Camerer C, Montague PR. A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci.* (2008) **9**:545–56. doi: 10.1038/nrn2357

135. Etkin A, Egner T, Kalisch R. Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends Cogn Sci.* (2010) **15**:85–93. doi: 10.1016/j.tics.2010.11.004

136. Lagioia A, Van De Ville D, Debbané M, Lazeyras F, Eliez S. Adolescent resting state networks and their associations with schizotypal trait expression. *Front Syst Neurosci.* (2010) **4**:35. doi: 10.3389/fnsys.2010.00035

137. Mantini D, Perrucci MG, Del Gratta C, Romani GL, Corbetta M. Electrophysiological signatures of resting state networks in the human brain. *Proc Natl Acad Sci USA* (2007) **104**:13170–13175. doi: 10.1073/pnas.0700668104

138. Ganis G, Thompson WL, Kosslyn SM. Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cogn Brain Res.* (2004) **20**:226–41. doi: 10.1016/j.cogbrainres.2004.02.012

139. Schiltz C, Sorger B, Caldara R, Ahmed F, Mayer E, Goebel R, et al. Impaired face discrimination in acquired prosopagnosia is associated with abnormal response to individual faces in the right middle fusiform gyrus. *Cereb. Cortex* (2006) **16**:574–586. doi: 10.1093/cercor/bhj005

140. Renier LA, Anurova I, De Volder AG, Carlson S, VanMeter J, Rauschecker JP. Preserved functional specialization for spatial processing in the middle occipital gyrus of the early blind. *Neuron* (2010) **68**:138–48. doi: 10.1016/j.neuron.2010.09.021

141. Vanni S, Tanskanen T, Seppä M, Uutela K, Hari R. Coinciding early activation of the human primary visual cortex and anteromedial cuneus. *Proc Natl Acad Sci USA* (2001) **98**:2776–780. doi: 10.1073/pnas.041600898

142. Kandel ER, Schwartz JH, Jessell TM. Principles of neural science, *Neurology* (2000). doi: 10.1036/0838577016

143. Kornhuber HH. Cortex, basal ganglia and cerebellum in motor control. *Electroencephalogr Clin Neurophysiol Suppl.* (1978) **34**:449–455.

144. Hikosaka O, Takikawa Y, Kawagoe R. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol Rev.* (2000) **80**:953–78. doi: 10.1152/physrev.2000.80.3.953

145. Stocco A, Lebiere C, Anderson JR. Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychol Rev.* (2010) **117**:541–74. doi: 10.1037/a0019077

146. Bennett M, Dennett D, Hacker P, Searle J, Hurford JR, Bennett M, et al. Neuroscience and philosophy: brain, mind, and language. *Q Rev Biol.* (2007) **82**:439–40. doi: 10.1086/527640

147. Soares JM, Sampaio A, Ferreira LM, Santos NC, Marques P, Marques F, et al. Stress impact on resting state brain networks. *PLoS ONE* (2013) **8**:e66500. doi: 10.1371/journal.pone.0066500

148. Habas C, Kamdar N, Nguyen D, Prater K, Beckmann CF, Menon V, et al. Distinct cerebellar contributions to intrinsic connectivity networks. *J Neurosci.* (2009) **29**:8586–94. doi: 10.1523/JNEUROSCI.1868-09.2009

149. Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc Natl Acad Sci USA* (2003) **100**:253–8. doi: 10.1073/pnas.0135058100

150. Chenji S, Jha S, Lee D, Brown M, Seres P, Mah D, et al. Investigating default mode and sensorimotor network connectivity in amyotrophic lateral sclerosis. *PLoS ONE* (2016) **11**:e0157443. doi: 10.1371/journal.pone.0157443

151. Friederici AD. The brain basis of language processing: from structure to function. *Physiol Rev.* (2011) **91**:1357–92. doi: 10.1152/physrev.00006.2011

152. Friederici AD, Gierhan SME. The language network. *Curr Opin Neurobiol.* (2013) **23**:250–54. doi: 10.1016/j.conb.2012.10.002

153. Emmorey K. The role of Broca's area in sign language. In: Grodzinsky Y, Amunts K, editors. *Broca's Region*. Oxford, UK: Oxford University Press (2006). pp. 167–82.

154. Mason RA, Prat CS, Just MA. Neurocognitive brain response to transient impairment of wernicke's area. *Cereb Cortex* (2014) **24**:1474–84. doi: 10.1093/cercor/bhs423

155. Acheson DJ, Hagoort P. Stimulating the brain's language network: syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. *J Cogn Neurosci.* (2013) **25**:1664–77. doi: 10.1162/jocn_a_00430

156. Fox MD, Corbetta M, Snyder AZ, Vincent JL, Raichle ME. Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proc Natl Acad Sci USA* (2006) **103**:10046–51. doi: 10.1073/pnas.0604187103

157. Yantis S. Goal-directed and stimulus-driven determinants of attentional control. *Control Cogn. Process. Atten. Perform.* (2000) **Xviii**:73–103. doi: 10.2337/db11-0571

158. Duan X, Liao W, Liang D, Qiu L, Gao Q, Liu C, et al. Large-scale brain networks in board game experts: insights from a domain-related task and task-free resting state. *PLoS ONE* (2012) **7**:e32532. doi: 10.1371/journal.pone.0032532

159. Majerus S, Attout L, D'Argembeau A, Degueldre C, Fias W, Maquet P, et al. Attention supports verbal short-term memory via competition between dorsal and ventral attention networks. *Cereb Cortex* (2012) **22**:1086–97. doi: 10.1093/cercor/bhr174

160. Fox MD, Raichle ME. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci.* (2007) **8**:700–11. doi: 10.1038/nrn2201

161. Shaw P, Lerch J, Greenstein D, Sharp W, Clasen L, Evans A, et al. Longitudinal mapping of cortical thickness and clinical outcome in children and adolescents with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* (2006) **63**:540–9. doi: 10.1001/archpsyc.63.5.540

162. Cubillo A, Halari R, Ecker C, Giampietro V, Taylor E, Rubia K. Reduced activation and inter-regional functional connectivity of fronto-striatal networks in adults with childhood Attention-Deficit Hyperactivity Disorder (ADHD) and persisting symptoms during tasks of motor inhibition and cognitive switching. *J Psychiatr Res.* (2010) **44**:629–39. doi: 10.1016/j.jpsychires.2009.11.016

163. Rubia K, Halari R, Cubillo A, Mohammad AM, Scott S, Brammer M. Disorder-specific inferior prefrontal hypofunction in boys with pure attention-deficit/hyperactivity disorder compared to boys with pure conduct disorder during cognitive flexibility. *Hum Brain Mapp.* (2010) **31**:1823–33. doi: 10.1002/hbm.20975

164. Liddle EB, Hollis C, Batty MJ, Groom MJ, Totman JJ, Liotti M, et al. Task-related default mode network modulation and inhibitory control in ADHD: effects of motivation and methylphenidate. *J Child Psychol Psychiatry.* (2011) **52**:761–71. doi: 10.1111/j.1469-7610.2010.02333.x

165. Konrad K, Eickhoff SB. Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder. *Hum Brain Mapp.* (2010) **31**:904–16. doi: 10.1002/hbm.21058

166. Li X, Sroubek A, Kelly MS, Lesser I, Sussman E, He Y, et al. Atypical pulvinar-cortical pathways during sustained attention performance in children with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry* (2012) **51**:1197–1207.e4. doi: 10.1016/j.jaac.2012.08.013

167. Cao X, Cao Q, Long X, Sun L, Sui M, Zhu C, et al. Abnormal resting-state functional connectivity patterns of the putamen in medication-naive children with attention deficit hyperactivity disorder. *Brain Res.* (2009) **1303**:195–206. doi: 10.1016/j.brainres.2009.08.029

168. Bush G. Attention-deficit/hyperactivity disorder and attention networks. *Neuropsychopharmacology* (2010) **35**:278–300. doi: 10.1038/npp.2009.120

169. Castellanos FX, Margulies DS, Kelly C, Uddin LQ, Ghaffari M, Kirsch A, et al. Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. *Biol. Psychiatry* (2008) **63**:332–7. doi: 10.1016/j.biopsych.2007.06.025

170. Sun L, Cao Q, Long X, Sui M, Cao X, Zhu C, et al. Abnormal functional connectivity between the anterior cingulate and the default mode network in drug-naïve boys with attention deficit hyperactivity disorder. *Psychiatry Res.* (2012) **201**:120–7. doi: 10.1016/j.pscychresns.2011.07.001

171. Bush G. Cingulate, frontal, and parietal cortical dysfunction in attention-deficit/hyperactivity disorder. *Biol Psychiatry* (2011) **69**:1160–7. doi: 10.1016/j.biopsych.2011.01.022

172. Castellanos FX, Proal E. Large-scale brain systems in ADHD: beyond the prefrontal-striatal model. *Trends Cogn Sci.* (2012) **16**:17–26. doi: 10.1016/j.tics.2011.11.007

173. Cubillo A, Rubia K. Structural and functional brain imaging in adult attention-deficit/hyperactivity disorder. *Expert Rev Neurother.* (2010) **10**:603–20. doi: 10.1586/ern.10.4

174. Köchel A, Schöngaßner F, Feierl-Gsodam S, and Schienle A. Processing of affective prosody in boys suffering from attention deficit hyperactivity disorder: a near-infrared spectroscopy study. *Soc Neurosci.* (2015) **919**:1–9. doi: 10.1080/17470919.2015.1017111

175. Kessler DA, Angstadt M, Welsh RC, Sripada, C. Modality-spanning deficits in attention-deficit/hyperactivity disorder in functional networks, gray matter, and white matter. *J Neurosci.* (2014) **34**:16555–66. doi: 10.1523/JNEUROSCI.3156-14.2014

176. Dickstein SG, Bannon K, Xavier Castellanos F, Milham MP. The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis. *J Child Psychol Psychiatry Allied Discip.* (2006) **47**:1051–62. doi: 10.1111/j.1469-7610.2006.01671.x

177. Rubia K. "Cool" inferior frontostriatal dysfunction in attention-deficit/hyperactivity disorder versus "hot" ventromedial orbitofrontal-limbic dysfunction in conduct disorder: a review. *Biol Psychiatry* (2011) **69**:e69–87. doi: 10.1016/j.biopsych.2010.09.023

178. Tian L, Jiang T, Wang Y, Zang Y, He Y, Liang M, et al. Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neurosci Lett.* (2006) **400**:39–43. doi: 10.1016/j.neulet.2006.02.022

179. Seidman LJ, Valera EM, Makris N. Structural brain imaging of attention-deficit/hyperactivity disorder. *Biol Psychiatry* (2005) **57**:1263–72. doi: 10.1016/j.biopsych.2004.11.019

180. Hale TS, Kane AM, Kaminsky O, Tung KL, Wiley JF, McGough JJ, et al. Visual network asymmetry and default mode network function in ADHD: An fMRI study. *Front. Psychiatry* (2014) **5**:81. doi: 10.3389/fpsyt.2014.00081

181. Ewers M, Sperling RA, Klunk WE, Weiner MW, Hampel H. Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. *Trends Neurosci.* (2011) **34**:430–442. doi: 10.1016/j.tins.2011.05.005

182. Greicius MD, Srivastava G, Reiss AL, Menon V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl Acad Sci USA* (2004) **101**:4637–42. doi: 10.1073/pnas.0308627101

183. Supekar K, Menon V, Rubin D, Musen M, Greicius MD. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Comput Biol.* (2008) **4**:e1000100. doi: 10.1371/journal.pcbi.1000100

184. Wang L, Zang Y, He Y, Liang M, Zhang X, Tian L, et al. Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *Neuroimage* (2006) **31**:496–504. doi: 10.1016/j.neuroimage.2005.12.033

185. Zhou J, Greicius MD, Gennatas ED, Growdon ME, Jang JY, Rabinovici GD, et al. Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. *Brain* (2010) **133**:1352–67. doi: 10.1093/brain/awq075

186. Gomez-Ramirez J, Wu J. Network-based biomarkers in Alzheimer's disease: review and future directions. *Front Aging Neurosci.* (2014) **6**:1–9. doi: 10.3389/fnagi.2014.00012

187. Buckner RL, Sepulcre J, Talukdar T, Krienen FM, Liu H, Hedden T, et al. a Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *J Neurosci.* (2009) **29**:1860–73. doi: 10.1523/JNEUROSCI.5062-08.2009

188. Sperling RA, Dickerson BC, Pihlajamaki M, Vannini P, LaViolette PS, Vitolo OV, et al. Functional alterations in memory networks in early Alzheimer's disease. *Neuromolecular Med.* (2010) **12**:27–43. doi: 10.1007/s12017-009-8109-7

189. Machulda MM, Ward HA, Borowski B, Gunter JL, Cha RH, O'Brien PC, et al. Comparison of memory fMRI response among normal, MCI, and Alzheimer's patients. *Neurology* (2003) **61**:500–6. doi: 10.1055/s-0029-1237430.Imprinting

190. Johnson SC, Schmitz TW, Trivedi M, a, Ries ML, Torgerson BM, Carlsson CM, et al. The influence of Alzheimer disease family history and apolipoprotein E epsilon4 on mesial temporal lobe activation. *J Neurosci.* (2006) **26**:6069–76. doi: 10.1523/JNEUROSCI.0959-06.2006

191. Agosta F, Pievani M, Geroldi C, Copetti M, Frisoni GB, Filippi M. Resting state fMRI in Alzheimer's disease: beyond the default mode network. *Neurobiol Aging* (2012) **33**:1564–78. doi: 10.1016/j.neurobiolaging.2011.06.007

192. Foundas AL, Leonard CM, Mahoney SM, Agee OF, Heilman KM. Atrophy of the hippocampus, parietal cortex, and insula in Alzheimer's disease: a volumetric magnetic resonance imaging study. *Neuropsychiatry Neuropsychol Behav Neurol.* (1997) **10**:81–9.

193. Rombouts SARB, Barkhof F, Witter MP, Scheltens P. Unbiased whole-brain analysis of gray matter loss in Alzheimer's disease. *Neurosci Lett.* (2000) **285**:231–3. doi: 10.1016/S0304-3940(00)01067-3

194. Karas GB, Scheltens P, Rombouts SARB, Visser PJ, Van Schijndel RA, Fox NC, et al. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *Neuroimage* (2004) **23**:708–16. doi: 10.1016/j.neuroimage.2004.07.006

195. Li R, Wu X, Chen K, Fleisher AS, Reiman EM, Yao L. Alterations of directional connectivity among resting-state networks in Alzheimer disease. *Am J Neuroradiol.* (2013) **34**:340–5. doi: 10.3174/ajnr.A3197

196. Monk CS, Peltier SJ, Wiggins JL, Weng, SJ, Carrasco M, Risi S, et al. Abnormalities of intrinsic functional connectivity in autism spectrum disorders. *Neuroimage* (2009) **47**:764–72. doi: 10.1016/j.neuroimage.2009.04.069

197. Assaf M, Jagannathan K, Calhoun VD, Miller L, Stevens MC, Sahl R, et al. Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. *Neuroimage* (2010) **53**:247–56. doi: 10.1016/j.neuroimage.2010.05.067

198. Radua J, Via E, Catani M, Mataix-Cols D. Voxel-based meta-analysis of regional white-matter volume differences in autism spectrum disorder versus healthy controls. *Psychol Med.* (2011) **41**:1539–50. doi: 10.1017/S0033291710002187

199. Martino A, Ross K, Uddin LQ, Sklar AB, Castellanos FX, and Milham MP. Functional brain correlates of social and nonsocial processes in autism spectrum disorders: an activation likelihood estimation meta-analysis. *Biol Psychiatry* (2009) **65**:63–74. doi: 10.1016/j.biopsych.2008.09.022

200. Weng SJ, Wiggins JL, Peltier SJ, Carrasco M, Risi S, Lord C, et al. Alterations of resting state functional connectivity in the default network in adolescents with autism spectrum disorders. *Brain Res.* (2010) **1313**:202–14. doi: 10.1016/j.brainres.2009.11.057

201. Mody M, Manoach DS, Guenther FH, Kenet T, Bruno KA, Mcdougle CJ, et al. Speech and language in autism spectrum disorder: a view through the lens of behavior and brain imaging. *Neuropsychiatry (London).* (2013) **3**:223–32. doi: 10.2217/npy.13.19

202. Verly M, Verhoeven J, Zink I, Mantini D, Peeters R, Deprez S, et al. Altered functional connectivity of the language network in ASD: role of classical language areas and cerebellum. *NeuroImage Clin.* (2014) **4**:374–82. doi: 10.1016/j.nicl.2014.01.008

203. Peeva MG, Tourville JA, Agam Y, Holland B, Manoach DS, Guenther FH. White matter impairment in the speech network of individuals with autism spectrum disorder. *NeuroImage Clin.* (2013) **3**:234–41. doi: 10.1016/j.nicl.2013.08.011

204. Prat CS, Stocco A, Neuhaus E, Kleinhans NM. Basal ganglia impairments in autism spectrum disorder are related to abnormal signal gating to prefrontal cortex. *Neuropsychologia* (2016) **91**:268–81. doi: 10.1016/j.neuropsychologia.2016.08.007

205. Uddin LQ, Menon V. The anterior insula in autism: under-connected and under-examined. *Neurosci Biobehav Rev.* (2009) **33**:1198–203. doi: 10.1016/j.neubiorev.2009.06.002

206. Uddin LQ, Supekar K, Lynch CJ, Khouzam A, Phillips J, Feinstein C, et al. Salience network-based classification and prediction of symptom severity in children with autism. *JAMA Psychiatry* (2013) **70**:869–79. doi: 10.1001/jamapsychiatry.2013.104

207. Birn RM, Patriat R, Phillips ML, Germain A, Herringa RJ. Childhood maltreatment and combat posttraumatic stress differentially predict fear-related fronto-subcortical connectivity. *Depress Anxiety* (2014) 31:880–92. doi: 10.1002/da.22291

208. Chen AC, Etkin A. Hippocampal network connectivity and activation differentiates post-traumatic stress disorder from generalized anxiety disorder. *Neuropsychopharmacology* (2013) 38:1889–98. doi: 10.1038/npp.2013.122

209. Cisler JM, Steele JS, Lenow JK, Smitherman S, Everett B, Messias E, et al. Functional reorganization of neural networks during repeated exposure to the traumatic memory in posttraumatic stress disorder: an exploratory fMRI study. *J Psychiatr Res.* (2014) 48:47–55. doi: 10.1016/j.jpsychires.2013.09.013

210. Braun K. The Prefrontal-limbic system: development, neuroanatomy, function, and implications for socioemotional development. *Clin Perinatol.* (2011) 38:685–702. doi: 10.1016/j.clp.2011.08.013

211. Bluhm RL, Williamson PC, Osuch EA, Frewen PA, Stevens TK, Boksman K, et al. Alterations in default network connectivity in posttraumatic stress disorder related to early-life trauma. *J Psychiatry Neurosci.* (2009) 34:187–94.

212. Long Z, Duan X, Xie B, Du H, Li R, Xu Q, et al. Altered brain structural connectivity in post-traumatic stress disorder: a diffusion tensor imaging tractography study. *J Affect Disord.* (2013) 150:798–806. doi: 10.1016/j.jad.2013.03.004

213. White SF, Costanzo ME, Blair JR, Roy MJ. PTSD symptom severity is associated with increased recruitment of top-down attentional control in a trauma-exposed sample. *NeuroImage Clin.* (2015) 7:19–27. doi: 10.1016/j.nicl.2014.11.012

214. Rabinak CA, Angstadt M, Welsh RC, Kenndy AE, Lyubkin M, Martis B, et al. Altered amygdala resting-state functional connectivity in post-traumatic stress disorder. *Front Psychiatry* (2011) 2:1–8. doi: 10.3389/fpsyt.2011.00062

215. Ebdlahad S, Nofzinger EA, James JA, Buysse DJ, Price JC, and Germain A. Comparing neural correlates of REM sleep in posttraumatic stress disorder and depression: a neuroimaging study. *Psychiatry Res.* (2013) 214:422–8. doi: 10.1016/j.pscychresns.2013.09.007

216. Germain A, James J, Insana S, Herringa RJ, Mammen O, Price J, et al. A window into the invisible wound of war: Functional neuroimaging of REM sleep in returning combat veterans with PTSD. *Psychiatry Res.* (2013) 211:176–9. doi: 10.1016/j.pscychresns.2012.05.007

217. Lanius RA, Bluhm R, Lanius U, and Pain C. A review of neuroimaging studies in PTSD: Heterogeneity of response to symptom provocation. *J Psychiatr Res.* (2006) 40:709–29. doi: 10.1016/j.jpsychires.2005.07.007

218. Stark EA, Parsons CE, Van Hartevelt TJ, Charquero-Ballester M, McManners H, Ehlers A, et al. Post-traumatic stress influences the brain even in the absence of symptoms: a systematic, quantitative meta-analysis of neuroimaging studies. *Neurosci Biobehav Rev.* (2015) 56:207–21. doi: 10.1016/j.neubiorev.2015.07.007

219. Herringa R, Phillips M, Almeida J, Insana S, Germain A. Post-traumatic stress symptoms correlate with smaller subgenual cingulate, caudate, and insula volumes in unmedicated combat veterans. *Psychiatry Res.* (2012) 203:139–45. doi: 10.1016/j.pscychresns.2012.02.005

220. Yin Y, Jin C, Eyler LT, Jin H, Hu X, Duan L, et al. Altered regional homogeneity in post-traumatic stress disorder: a restingstate functional magnetic resonance imaging study. *Neurosci Bull.* (2012) 28:541–9. doi: 10.1007/s12264-012-1261-3

221. Faries DE, Yalcin I, Harder D, Heiligenstein JH. Validation of the ADHD Rating Scale as a clirician administered and scored instrument. *J Atten Disord.* (2001) 5:107–15. doi: 10.1177/108705470100500204

222. Zhang S, Faries DE, Vowles M, Michelson D. ADHD Rating Scale IV: psychometric properties from a multinational study as a clinician-administered instrument. *Int J Methods Psychiatr Res.* (2005) 14:186–201. doi: 10.1002/mpr.7

223. Thaler NS, Bello DT, Etcoff LM. WISC-IV profiles are associated with differences in symptomatology and outcome in children with ADHD. *J Atten Disord.* (2013) 17:291–301. doi: 10.1177/1087054711428806

224. Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol.* (2013) 9:106–18. doi: 10.1038/nrneurol.2012.263

225. Morris JC. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *Int Psychogeriatr.* (1997) (9 Suppl 1):173–176–178. doi: 10.1017/S1041610297004870

226. Williams MM, Storandt M, Roe CM, Morris JC. Progression of Alzheimer's disease as measured by clinical dementia rating sum of boxes scores. *Alzheimers Dement.* (2013) 9:S39–44. doi: 10.1016/j.jalz.2012.01.005

227. Teng E, Becker BW, Woo E, Knopman DS, Cummings JL, Lu PH. Utility of the functional activities questionnaire for distinguishing mild cognitive impairment from very mild alzheimer disease. *Alzheimer Dis Assoc Disord.* (2010) 24:348–53. doi: 10.1097/wad.0b013e3181e2fc84

228. Bastiaansen JA, Meffert H, Hein S, Huizinga P, Ketelaars C, Pijnenborg M, et al. Diagnosing autism spectrum disorders in adults: the use of Autism Diagnostic Observation Schedule (ADOS) module (2011) 4. *J Autism Dev Disord.* 41:1256–66. doi: 10.1007/s10803-010-1157-x

229. Gotham K, Risi S, Pickles A, Lord C. The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. *J Autism Dev Disord.* (2007) 37:613–27. doi: 10.1007/s10803-006-0280-1

230. Siegel DJ, Minshew NJ, Goldstein G. Wechsler IQ profiles in diagnosis of high-functioning autism. *J Autism Dev Disord.* (1996) 26:389–406. doi: 10.1007/BF02172825

231. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, et al. Austism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord.* (1989) 19:185–212. doi: 10.1007/BF02211841

232. Bryant RA. Early predictors of posttraumatic stress disorder. *Biol Psychiatry* (2003) 53:789–95. doi: 10.1016/S0006-3223(02)01895-4

233. Burriss L, Ayers E, Ginsberg J, Powell DA. Learning and memory impairment in PTSD: relationship to depression. *Depress Anxiety* (2008) 25:149–57. doi: 10.1002/da.20291

234. Dretsch MN, Thiel KJ, Athy JR, Born S, Prue-Owens K. Posttraumatic stress disorder in the U.S. Warfighter: sensitivity to punishment and antidepressant use contribute to decision-making performance. *Traumatology* (2013) 19:118–25. doi: 10.1177/1534765612455228

235. Keane TM, Fairbank JA, Caddell JM, Zimering RT, Taylor KL, Mora CA. Clinical evaluation of a measure to assess combat exposure. *Psychol. Assess.* (1989) 1:53–55. doi: 10.1037//1040-3590.1.1.53

236. Koenen KC, Lyons MJ, Goldberg J, Simpson J, Williams WM, Toomey R, et al. Co-twin control study of relationships among combat exposure, combat-related PTSD, and other mental disorders. *J Trauma Stress* (2003) 16:433–8. doi: 10.1023/A:1025786925483

237. Britt TW, Adler AB, Bliese PD, Moore D. Morale as a moderator of the combat exposure-PTSD symptom relationship. *J Trauma Stress* (2013) 26:94–101. doi: 10.1002/jts.21775

238. Gray MJ, Litz BT, Hsu JL, Lombardo TW. Psychometric properties of the life events checklist. *Assessment* (2004) 11:330–41. doi: 10.1177/1073191104269954

239. Zhao X, Rangaprakash D Jr, Katz JS, Dretsch MN, Deshpande G. Data and code for identifying different neuropsychiatric disorders using unsupervised clustering methods. *Data Brief* (2018). doi: 10.1016/j.dib.2018.01.080

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership