RESEARCH METHODS PEDAGOGY: ENGAGING PSYCHOLOGY STUDENTS IN RESEARCH METHODS AND STATISTICS

EDITED BY : Lynne D. Roberts

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**

RESEARCH METHODS PEDAGOGY: ENGAGING PSYCHOLOGY STUDENTS IN RESEARCH METHODS AND STATISTICS

Topic Editor:
Lynne D. Roberts, Curtin University, Australia

Research methods and statistics are central to the development of professional competence and evidence based psychological practice. Furthermore, the ability to interpret and apply research findings contributes to the development of psychological literacy, the primary outcome of an undergraduate education in psychology. Despite this, many psychology students express little interest in, and in some cases an active dislike of, learning research methods and statistics. This ebook brings together current research, innovative evidence-based practice and critical discourse related to engaging psychology students in learning quantitative and qualitive research methods.

Table of Contents

frontiers
in Psychology

Check for
updates

Editorial: Research Methods Pedagogy: Engaging Psychology Students in Research Methods and Statistics

*Lynne D. Roberts** *

School of Psychology and Speech Pathology, Curtin University, Perth, WA, Australia

The Editorial on the Research Topic

Research Methods Pedagogy: Engaging Psychology Students in Research Methods and Statistics

Across disciplines, many introductory research methods students present as uninterested and unmotivated to learn a topic they see as lacking in relevance (Earley, 2014); psychology is no exception (Ruggeri et al., 2008). This is problematic as research methods and statistics are central to the development of professional competence and evidence based psychological practice. Reflecting this, the ability to interpret, design, and conduct basic psychological research forms part of the "scientific inquiry and critical thinking" goal of the *APA Guidelines for the Undergraduate Psychology Major: Version 2.0* (American Psychological Association, 2016), with research methods and statistics a requirement of almost all undergraduate psychology programs (Norcross et al., 2016). Furthermore, the ability to interpret and apply research findings contributes to the development of psychological literacy, the primary outcome of an undergraduate education in psychology (Cranney and Dunn, 2011). This Research Topic brings together current research, innovative evidence-based practice and critical discourse related to engaging undergraduate psychology students in learning quantitative and qualitative methods research.

In the first of fourteen articles, Lacot et al. present a perspective on two methods of stimulating first year undergraduate students' interest in research methods within an introductory psychology course, with the aim of promoting critical thinking about research right from the beginning of the undergraduate degree. Teaching students to critically question what they hear and read, to "evidence check" and propose ways of testing are important components underlying the development of research methods competence.

While interest in research methods can be stimulated in introductory psychology course, most teaching of research methods occurs within dedicated research methods courses. Three articles explore the role of active learning within research methods courses in psychology, a strategy with demonstrated effectiveness for increasing student performance across science, technology, engineering and mathematics courses (Freeman et al., 2014). Allen and Baughman report that in comparison to students in didactic workshops, students in activity based workshops demonstrated greater knowledge of, and confidence in using, research methods, but did not differ in their satisfaction with the learning experience. Rock et al. describe how eLearning systems might be used to actively engage psychology students in research methods and statistics through the application of eLearning pedagogical principles, providing examples of teaching advanced research methods within a virtual world. Lim et al. report that retrieval practice produces better long-term retention of statistical knowledge than does repeated studying. These articles share a focus on the benefits of

actively engaging psychology undergraduate students in learning research methods and statistics.

Structural knowledge of research methods can be developed through formal methodology training (Balloo et al., 2016). In an original research article Allen et al. identify a key area of difficulty for psychology students is selecting appropriate statistical tests for analysing data. Provided with a range of vignettes depicting common research scenarios, undergraduate psychology students struggled to articulate a systematic decision making process for selecting statistical tests. This finding highlights the need for a statistical decision making aid based on decision-tree logic to support student learning. Building on this, Allen et al.'s technology report introduces a free mobile app, (Allen et al., 2015), designed to do just that. StatHand guides students through a systematic process for identifying an appropriate statistical test for a wide range of research scenarios, scaffolding learning through a focus on the structural features of research scenarios.

Two further articles focus on the use of technology to support student learning of research methods and statistics. Ellis and Merdian highlight the use of dynamic and interactive data visualizations using the open-source statistical software *R* with a range of packages, such as (Chang et al., 2015), designed for this purpose. These packages enable teachers and students to modify visualizations in real time through, for example, selecting subsets of data to examine and compare, or manipulating inputs. Dynamic, interactive data visualizations have the potential to replace the static graphical illustrations commonly used in teaching research methods statistics. Moreau provides a list of recommended online resources that can be accessed by academics wishing to incorporate dynamic integrative data visualizations into their own teaching of research methods to enhance student learning.

A series of three articles by Perezgonzalez focus on the Null Hypothesis Significant Testing (NHST) controversy as it applies to the teaching of statistics in psychology. Once described as "surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" (Rozeboom, 1997, p. 335), NHST remains a commonly used analytic approach in psychology, although now with requirements for the addition of effect sizes, confidence intervals and descriptive text (American Psychological Association, 2009). In the first opinion piece, Perezgonzalez addresses theoretical misinterpretations regarding statistical significance. This is followed by a general commentary article on the use of

p-values, where he illustrates the differences in interpretation depending upon whether a percentile or probability heuristic is used. The third review article presents a tutorial for teaching hypothesis testing theories. Working from a different perspective Aksentijevic presents statistics anxiety in psychology students as a rational response to myths about the nature of probability and statistics. In his perspective piece Aksentijevic suggests the focus needs to shift away from NHST to the larger role of statistics in research.

Quantitative research methods and statistics predominate in the teaching of research methods within psychology. Perhaps reflecting this, only one article in this research topic focused on teaching qualitative methods. Roberts and Castell examine third year undergraduate psychology students' attitudes to qualitative research. Students viewed qualitative research as a paradigmatic shift requiring new ways of thinking about research that was alternatively construed as a threat or an advantage. Roberts and Castell advocate for the integration of teaching of qualitative and quantitative research methods to reduce the perceived dichotomy between the two.

Rounding off this special issue is an opinion piece on scientific integrity. Schoenherr argues for the explicit inclusion of scientific integrity in the undergraduate psychology curriculum. The importance of this is highlighted by research identifying research misconduct and questionable research practices by psychology students (Rajah-Kanagasabai and Roberts, 2015).

The articles in this research topic contribute to what has been referred to as the "under-researched and under-developed" pedagogical culture for teaching research methods in the social sciences (Lewthwaite and Nind, 2016). The combination of strategies, practices, and recommended software provides psychology academics with new avenues for teaching research methods and supporting student learning. As found by Allen and Baughman, and reported previously by Sizemore and Lewandowsky (2009), better outcomes do not necessarily equate with more positive attitudes toward research methods. These findings highlight the importance of addressing attitudes in addition to increasing students' knowledge and application of research methods.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Allen, P., Roberts, L., Baughman, F., van Rooy, D., Rock, A., and Loxton, N. (2015). *StatHand [computer software]*. Sydney: Office for Learning and Teaching. Available online at: https://www.stathand.net/

American Psychological Association (2009). *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

American Psychological Association (2016). Guidelines for the undergraduate psychology major: Version 2.0. *Am. Psychol.* 71, 102–111. doi: 10.1037/a0037562

Balloo, K., Pauli, R., and Worrell, M. (2016). Individual differences in psychology undergraduates' development of research methods knowledge and skills. *Procedia Soc. Behav. Sci.* 217, 790–800. doi: 10.1016/j.sbspro.2016.02.147

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2015). *Shiny: Web Application Framework for R*. R package version 0.11, 1.

Cranney, J., and Dunn, D. S. (2011). *The Psychologically Literate Citizen: Foundations and Global Perspectives*. Oxford: Oxford University Press.

Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teach. High. Educ.* 19, 242–253. doi: 10.1080/13562517.2013.860105

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science,

engineering, and mathematics. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8410–8415. doi: 10.1073/pnas.1319030111

Lewthwaite, S., and Nind, M. (2016). Teaching research methods in the social sciences: expert perspectives on pedagogy and practice. *Br. J. Educ. Stud.* 1–17. doi: 10.1080/00071005.2016.1197882

Norcross, J. C., Hailstorks, R., Aiken, L. S., and Pfund, R. A. (2016). Undergraduate study in psychology: curriculum and assessment. *Am. Psychol.* 71, 89–101. doi: 10.1037/a0040095

Rajah-Kanagasabai, C. J., and Roberts, L. D. (2015). Predicting self-reported research misconduct and questionable research practices in university students using an augmented Theory of Planned Behavior. *Front. Psychol.* 6:535. doi: 10.3389/fpsyg.2015.00535

Rozeboom, W. W. (1997). "Good science is abductive, not hypothetico-deductive," in *What if there were no Significance Tests?* eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah, NJ: Erlbaum), 335–392.

Ruggeri, K., Dempster, M., Hanna, D., and Cleary, C. (2008). Experiences and expectations: the real reason nobody likes stats. *Psychol. Teach. Rev.* 14, 74–83.

Sizemore, O. J., and Lewandowsky, G. W. (2009). Learning might not equal liking: research methods course changes knowledge but not attitudes. *Teach. Psychol.* 36, 90–95. doi: 10.1080/00986280902739727

From Bill Shankly to the Huffington Post: How to Increase Critical Thinking in Experimental Psychology Course?

Emilie Lacot[1,2], Geoffrey Blondelle[1] and Mathieu Hainselin[1]*

[1] *CRP-CPO, EA 7273, Université de Picardie Jules Verne, Amiens, France,* [2] *Centre d'Activité de Génétique Clinique et d'Oncogénétique, Centre Hospitalier Universitaire Amiens Picardie, Amiens, France*

Although critical thinking and source checking are basic prerequisites to become a psychologist, or a scientist, it is usually difficult to have students interested in experimental methods courses. Most first year students are tempted not to attend these courses. Such behaviors are reinforced by arguments that "everybody is different" and "people are not numbers." Consequently, students have difficulties to develop source and evidence checking skills, and may be more prone to believe in any supposed expert. This paper presents two ways to involve students during lectures and seminars. The first method consists in presenting, during the initial lecture of the year, a fake scientific concept which students will believe as true. This phenomenon is called the "Bill Shankly syndrome" and it only exists if someone believes that the information is given by a serious lecturer, presenting oneself as a world-class researcher. The second method consists in training students to become reviewers using evidence checking of a mainstream media article which promises scientifically proven ways to be happy. The use of these methods may stimulate students' interest in research methods and its practical applications from week one.

Keywords: critical thinking, pedagogy, belief, reviewing process, authority

INTRODUCTION

Psychology students, especially first year university students in France, attend their first lessons with many beliefs about psychology. Most of them think that it is not a scientific field, and are thus surprised with the number of courses on neuroscience, statistics and methodology. At the very beginning, students are confronted with the different steps of the scientific method and discover the mean of a sample, which is frequently and wrongly associated with the mean of a population. These concepts are usually in contradiction with their own conceptions such as "everybody is different" and "people are not numbers" (Malekoff, 2008). Due to these misconceptions, students may encounter difficulties in understanding or seeing the usefulness of such courses (Castro Sotos et al., 2007; Gigerenzer et al., 2007).

Moreover, for many reasons, it is hard for them to develop critical thinking or any experimental methodology skills. In high school, and more specifically in France, psychology is a tiny part of the Philosophy course, only studied through Freud's theories of psychoanalysis (Lieury, 2014). Although, APA have a resource manual for psychology

teaching including critical thinking (APA, 2015), students have barely been trained for critical thinking, as most part of their schooling is based on a passive listening model (e.g., Paul, 1992). Wegwert (2014, 141) highlighted that "fear is a powerful presence in schools, in teacher education, and in teacher identity" and consequently in students daily life. Because of the perceived power of lecturers, which can intimidate students, they feel compelled to integrate knowledge without checking its truthfulness. Consequently, students may be more prone to believe in any authority figure and are less inclined to show a critical mind (Wegwert, 2014). In the same vein, Reeve (2009), Skinner and Belmont (1993), highlight that the same phenomenon occurred if lecturers do not adopt an autonomy-supportive style. In our sense, these facts inhibit the development of great psychologists or scientists. However, a growing literature underlines that psychologists, scientists and thus lecturers can, due to a lack of evidence-based practice, make inappropriate inferences. For example, in their recent paper, Lilienfeld et al. (2014) highlight four barriers to scientific thinking: naïve realism, confirmation bias, illusory causation, and illusion of control. Thus, using concrete examples of inappropriate inferences in patients' treatment could help students identifying those biases. For example confirmation bias: I think I am a good therapist so patients have improved with me, when in fact it may come from an external event. The purpose is to show the many steps needed for any scientific reasoning, as well as for psychotherapies evaluation, and for possible clinical thinking.

We propose two different methods to engage first year psychology students and promote critical thinking. The first method, which must be used during the initial experimental methods lecture, consists of presenting a false scientific concept: the "Bill Shankly syndrome," presented in the first part of this article. This method is supposed to sharpen the critical mind of the student and to lead them to make their own scientific research. However, first year students of psychology tend to search for information on the Internet without verifying the scientific quality of the sources. To change this behavior, we have used, during seminars, a participative method based on the criticism of research supposedly defined as scientific. This second method, described in the second part of this article, allows the students to distinguish between a scientific source and a non-scientific source and also to criticize the methodology used.

THE BILL SHANKLY SYNDROME: A SERIOUS JOKE FOR A LESSON

We propose a method that was used with first year students of psychology at Jules Verne University of Picardy. This method consists in presenting a scientific concept: the "Bill Shankly syndrome" during the initial lecture of the year. However, the success of this method benefits from a specific context (i.e., the lecturer must have some authority). Before the presentation of the concept, the university lecturer introduces oneself to the students. The lecturer boasts very seriously about his career, asking students to call him "Doctor" (which is very unusual in France if you are not a Medical Doctor) due to his PhD,

explain that he is an Associate Professor at the Department of psychology, a trained neuropsychologist, works with high-level research teams in different countries, publishes articles and is asked for his expertise (i.e., as a reviewer) for international scientific journals. Although, this is a standard resume for an Associate Professor, students are unaware of it. Then, this peremptory introductory speech takes place in an unusually silent amphitheater. This assertive presentation is determinant in the Bill Shankly syndrome. Immediately, the PowerPoint lecture starts, with the classical pavlovian-writing behavior (students blindly copying after the presentation of each slide) and moving on to the next slide. Note that this type of presentation is important too. In fact, the use of PowerPoint animations enable a chronological presentation (i.e., scrolling the sentences one by one), better note taking and, as all lecturers hope, a better understanding of courses (Schmaltz and Enström, 2014). The concept that we arbitrarily called the Bill Shankly syndrome, because the last author (MH) is a Liverpool F.C. fan, is presented as a main concept in psychology. The lecturer expresses it as naturally and seriously as possible. It is written on slides and read out to the students that the syndrome consists in believing that any truth is the truth because this truth is named, expressed, illustrated as a scientific truth. This definition is followed by a reference to a fake scientific reference "Shankly et al., 1959" (see Supplementary Material for the slide), with a fake concept. The only real thing in this part is the Bill Shankly black and white picture. The definition remains deliberately vague to reinforce social influence (see below). Above all, choosing a 1960's Scottish football manager allow us to emphasize that anything, including old sports references, can be seen as scientific if students do not improve their critical thinking.

The concept that was introduced was clearly flawed, so that someone with critical thinking skills should question the validity of such a claim. To believe that everything is true because someone says so should be an aberration for psychologists, scientists or even for students. In principle, critical thinking involves questioning concepts and existing theories. However, our example highlights that the majority of the first year students agree with this concept. In the past couple of years, about 1000 students attended this course. They took notes without one single objection, and none of them asked any questions. This silence can be interpreted as the students' idea of university lecturers as having a great deal of knowledge. Note that these effects (i.e., silence and note taking without questioning) can be increased by informational or normative social influences and by conformity behaviors (students might assume the actions of others in an attempt to reflect correct behavior for a given situation). Indeed, if the majority of students write in silence without questioning courses, the others are more likely to do the same. It highlights the strong influence of peer group, the compliance and the conformity, particularly in a new situation with possible anxiety (Guimond, 1997; Cialdini and Goldstein, 2004). Altogether, this is a great place to discuss informational social influence with students.

Following the presentation of the Bill Shankly syndrome, the university lecturer explained that this concept is false. His speech was supported by a new sentence appearing on the slide: "the

TABLE 1 | Examples of tricks and justification to be happy from the studied article, with found references and possible criticizes.

Tricks	Justifications	References	Criticizes
1. Make a gift.	According *a study*, kind souls are happy.	Dunn et al., 2008	Scientific article in *Science*. Possible limits pinpointed because of financial resources of the participants.
2. Count his/her moments of happiness.	*A professor* at the University of Pennsylvania has demonstrated [...] were happier than the average.	Seligman, 2012	Book. Non-scientific. What does mean average? Compared to whom?
3. Live new experiences.	Nevertheless, *researches* that get out of the daily routine made happy.	MSN Lifestyle.	Non-scientific.
4. Anticipate good times.	None	Blog.	Non-scientific.
5. See the life in blue.	Sussex University scientists have shown that [...].	Interview of the concerned scientists retrieved from the website DailyMail, Macrae, 2009.	Non-scientific. Possibility of conflicts of interests.
6. Define targets.	[...] according to the psychologist Richard Davidson.	Jackson et al., 2000	Scientific article but it does not match with the tricks (different topic).
7. Stop always wanting to be right.	The writer and author Deepak Chopra recommends to his readers to remain neutral [...].	Interview of Deepak Chopra, FoxNews, 2011	Non-scientific. Does not match with the tricks.
8. Go to church.	[...] according to the Melbourne University.	DailyMail, 2010b	Non-scientific. The study is not published.
9. Sleep at least 6 hours a night.	In a study [...] assessed their level of happiness on a scale varying from 1 to 5 [...].	DailyMail, 2010a	Non-scientific. What does mean a level of happiness? Have we the same definition?
10. 20min in means of transport.	British scientists advise to limit the path that leads to work [...].	DailyMail, 2010a	Non-scientific. Some factors will be beyond the person's control.
11. Have at least 10 friends.	Authentic friend can be counted on fingers. This expression is familiar but for scientists [...].	Website DailyMail,Macrae, 2008	Non-scientific. How to define a friend as being a real friend?
12. Keep it positive even at wrong times.	According to a psychologist [...].	Website Howstuffworks, Layton, 2009	Non-scientific. Use of words as "it seems."
13. Don't forget be in love!	None.	None.	The entire trick!

Bill Shankly syndrome" obviously does not exist...unless you believe it...and thus you become a victim. Then the lecturer explained that William (Bill) Shankly (1913–1981) was a Scottish footballer and manager of Liverpool (Peace, 2014). During the time of these explanations, it was interesting to note that the majority of students were still writing down the PowerPoint's sentences. Nevertheless, some of them understood the joke lesson and then they initiated discussion about the lecturer's speech. Indeed, despite his position as a lecturer/expert, all content should be supported by expert scientific support. At this point, the fake concept was disclosed and the lecture started again from the beginning with an as-normal-as-possible presentation of the lecturer and the course. The pedagogical aim, besides the academic message, is to generate in students the feeling that they can be victims of several cognitive biases and more largely they can be victims of social phenomena. Indeed, due to the social status of the university lecturer, students believe his speech without questioning the situation or the contents. Thus the message given here is: "enhance your critical thinking: don't believe everything you listen or read whoever the speaker/author is." Our aim is that students keep an open and critical mind. This requires searching for scientific information outside the courses.

However, in the same way that students believe in the lecturer, they can also think that all Internet retrieved information is true and scientific, particularly if there is an expert cited (i.e., some students have cited blog posts as a scientific reference because the blogger claimed, incorrectly, to be an expert). The lecturer must warn them against false sources, and to check for original scientific source rather than just blindly believing indirect sources (i.e., self-proclaimed expert) instead of fact checking. During this first lecture, the experimental method, scientific journals standards, including peer-review process, are presented. As a follow up to this exercise, students are presented with an opportunity to seek out valid scientific sources.

FIRST REVIEWING EXPERIENCE: FIRST DISAPPOINTMENT

Like the first method presented, the second one is also used with the first year students of psychology in Jules Verne University of Picardy during seminars, always occurring after the first lecture with the Shankly effect. Thus, students are already aware of the importance of reconsidering the lecturer's discourse

content. They also know the need to verify supposedly scientific knowledge. Another important point is that this method must be used during seminars. As there are fewer students in seminars (taking place in classroom, 30–40) than in lectures (taking place in a lecture hall, 250–300), students are more likely to speak in public, which is usually a difficult exercise for them.

This second method consists in training students to become reviewers using evidence checking. More specifically, three points should be broached: (i) distinguish a scientific source (i.e., peer-reviewed journal) vs. a non-scientific source (i.e., without peer-review), (ii) criticize the proposed hypotheses, and (iii) propose another experiment to check, replicate or go further. In this way and after a brief presentation, the lecturer starts the PowerPoint course based on an article from the Le HuffPost (2013). It proposes 13 tricks scientifically proved to make people feel happy; each trick is displayed in **Table 1**. Tricks are seen one by one with the students.

First, only the trick and the original picture in the article were presented. Students were asked to say what they think about the trick and decide how they would validate the claim using source checking and critical thinking. For every trick, the first step is to find the source (scientific journal or not) by clicking on a link. The Huffington Post was chosen because it is a digital media allowing links, and because, except for this particular article, many good scientific popularization articles are available (Eustache, 2014). The aim is to make students check sources as often as possible, not to destroy any journal reputation.

When any limit or lack of scientific reference is underlined, they have to suggest a new experiment to assess the trick validity. This was a first step into scientific methods.

PERSPECTIVE

The Shankly effect and the media source checking are engaging exercises to teach experimental methods. After showing some surprise in the first place, students seem to like this new approach to enhance critical thinking (i.e., the Bill Shankly syndrome and its explanation), and use it beyond the specific course.

Although generalization is expected, there are cautions for instructors who are going to use this approach. The first concern is an over-generalization of the "don't believe everything you listen or read, whoever the speaker/author is" to every single point of every lesson. If too many students ask lots of justification questions during the lecture, it might prevent the lecturer from saying everything s/he would have liked to say. In our experience, it might also be passed on to other lessons and lecturers, for whom over-questioning and fact checking could be unusual. To avoid this, we ask for constructive criticism and always give our sources so students can check themselves. While these approaches have worked well for the authors, the evidence is anecdotal and as of yet, limited to students in France. Our methods are however consistent with recent recommendation to develop critical thinking (Schwanz and McIlreavy, 2015). We strongly encourage instructors to try these methods, as well as other engaging methods, to help promote critical thinking among students. Future research is needed to assess the actual efficiency, short and long term, of the Shankly effect.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct, and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00538

REFERENCES

APA (2015). *Resource Manual for New Teachers of High School Psychology.* Available onlineat: www.apa.org/ed/precollege/topss/resource-manual.aspx (Accessed March 1, 2016).

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., and Onghena, P. (2007). Students' misconceptions of statistical inference: a review of the empirical evidence from research on statistics education. *Educ. Res. Rev.* 2, 98–113. doi: 10.1016/j.edurev.2007.04.001

Cialdini, R. B., and Goldstein, N. J. (2004). Social Influence: compliance and conformity. *Annu. Rev. Psychol.* 55, 591–621. doi: 10.1146/annurev.psych.55.090902.142015

DailyMail (2010a). *Six Hours Sleep a Night, Four Shopping Sprees a Month and Two Holidays a Year among "Secrets to Happiness".* Available onliine at: http://www.dailymail.co.uk/news/article-1316165/Six-hours-sleep-night-shopping-sprees-month-holidays-year-secrets-happiness.html

DailyMail (2010b). *The Secret to Happiness: Don't Date a Neurotic or Worry About Your Career, Go to Church and Stay Thin.* Available online at: http://www.dailymail.co.uk/news/article-1318909/The-secret-happiness-dont-date-neurotic-worry-career-church-stay-thin.html

Dunn, E. W., Aknin, L. B., and Norton, M. I. (2008). Spending money on others promotes happiness. *Science* 319, 1687–1688. doi: 10.1126/science.1150952

Eustache, F. (2014). *A Partir de Quand Faut-il S'inquiéter Quand la Mémoire Nous Fait Défaut ? The Huffington Post.* Available online at: http://www.huffingtonpost.fr/francis-eustache/a-partir-de-quand-faut-il-sinquieter-quand-la-memoire-nous-fait-defaut_b_6307278.html

FoxNews (2011). *How to Make Yourself Happy - Right Now.* Available online at: http://www.foxnews.com/health/2011/05/31/make-happy-right.html

Gigerenzer, G., Gaissmaier, W., Kurz-milcke, E., Schwartz, L. M., and Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interes.* 8, 53–96. doi: 10.1111/j.1539-6053.2008.00033.x

Guimond, S. (1997). Attitude Change During College: Normative or Informational Social Influence? *Soc. Psychol. Educ.* 2, 237–261. doi: 10.1023/A:1009662807702

Jackson, D. C., Malmstadt, J. R., Larson, C. L., and Davidson, R. J. (2000). Suppression and enhancement of emotional responses to unpleasant pictures. *Psychophysiology* 37, 515–522. doi: 10.1111/1469-8986.3740515

Layton, J. (2009). *Evidence That Smiling Causes Happiness.* Available online at: http://science.howstuffworks.com/life/smiling-happy.htm

Le HuffPost (2013). *Comment se Sentir Heureux: 13 Astuces Prouvées Scientifiquement.* Available online at: http://www.huffingtonpost.fr/2013/07/05/sentir-heureux-astuces-prouvees-scientifiquement_n_3550465.html

Lieury, A. (2014). *Introduction à la Psychologie Cognitive.* Paris: Dunod.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., and Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: a taxonomy of causes of spurious therapeutic effectiveness. *Perspect. Psychol. Sci.* 9, 355–387. doi: 10.1177/1745691614535216

Macrae, F. (2008). *The Secret to Happiness? Having At Least 10 Good Friends.* Available online at: http://www.dailymail.co.uk/news/article-1079997/The-secret-happiness-Having-10-good-friends.html

Macrae, F. (2009). *Feeling Blue? Good News - Its the Colour of Happiness.* Available online at: http://www.dailymail.co.uk/sciencetech/article-1106236/Feeling-blue-Good-news--colour-happiness.html

Malekoff, A. (2008). A flexible organizing framework for group work with adolescents. *Soc. Work Groups* 30, 85–102. doi: 10.1300/J009v30n03_06

Paul, R. (1992). Critical thinking: what, why, and how. *New Dir. Community Coll.* 1992, 3–24. doi: 10.1002/cc.36819927703

Peace, D. (2014). *Red or Dead.* London: Faber & Faber Fiction.

Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educ. Psychol.* 44, 159–175. doi: 10.1080/00461520903028990

Schmaltz, R. M., and Enström, R. (2014). Death to weak PowerPoint: strategies to create effective visual presentations. *Front. Psychol.* 5:1138. doi: 10.3389/fpsyg.2014.01138

Schwanz, K., and McIlreavy, M. (2015). Academic performance of introductory psychology students: the importance of critical thinking. *Res. Psychol. Behav. Sci.* 3, 25–31. doi: 10.12691/rpbs-3-2-2

Seligman, M. E. P. (2012). *Flourish: A Visionary New Understanding of Happiness and Well-being.* Atria Books Available at: http://www.amazon.com/Flourish-Visionary-Understanding-Happiness-Well-being/dp/1439190763 (Accessed January 13, 2014).

Skinner, E. A., and Belmont, M. J. (1993). Motivation in the classroom: reciprocal effects of teacher behavior and student engagement across the school year. *J. Educ. Psychol.* 85, 571–581. doi: 10.1037/0022-0663.85.4.571

Wegwert, J. C. (2014). Looking the tiger in the eye: overcoming fear-based teacher identities. *Child. Educ.* 90, 137–142. doi: 10.1080/00094056.2014.894816

Active Learning in Research Methods Classes Is Associated with Higher Knowledge and Confidence, Though not Evaluations or Satisfaction

Peter J. Allen* and Frank D. Baughman

School of Psychology and Speech Pathology, Curtin University, Perth, WA, Australia

Research methods and statistics are regarded as difficult subjects to teach, fueling investigations into techniques that increase student engagement. Students enjoy active learning opportunities like hands-on demonstrations, authentic research participation, and working with real data. However, enhanced enjoyment does not always correspond with enhanced learning and performance. In this study, we developed a workshop activity in which students participated in a computer-based experiment and used class-generated data to run a range of statistical procedures. To enable evaluation, we developed a parallel, didactic/canned workshop, which was identical to the activity-based version, except that students were told about the experiment and used a pre-existing/canned dataset to perform their analyses. Tutorial groups were randomized to one of the two workshop versions, and 39 students completed a post-workshop evaluation questionnaire. A series of generalized linear mixed models suggested that, compared to the students in the didactic/canned condition, students exposed to the activity-based workshop displayed significantly greater knowledge of the methodological and statistical issues addressed in class, and were more confident about their ability to use this knowledge in the future. However, overall evaluations and satisfaction between the two groups were not reliably different. Implications of these findings and suggestions for future research are discussed.

Keywords: active learning, research methods, statistics, computer based experiments, authentic data, canned data

INTRODUCTION

A cornerstone of educational practice is the notion that the more engaged the learner, the more interested, passionate and motivated they will become, and the better the outcome will typically be vis-à-vis their learning. This causal chain, of sorts, thus predicts that higher rates of student retention, better grades, and higher levels of satisfaction and enjoyment are more likely to follow when a student is genuinely curious and involved in their study. However, student engagement appears to be more difficult to achieve in some areas of study compared to others. For instance, within psychology, research methods and statistics are widely regarded as 'difficult' subjects to teach (e.g., Conners et al., 1998). Student attitudes toward these topics are often negative (Murtonen, 2005; Sizemore and Lewandowski, 2009), and their interest in them is low (Vittengl et al., 2004; Rottinghaus et al., 2006). This lack of engagement is likely to impact

student outcomes, contributing to poorer grades and higher rates of attrition. However, a basic understanding of research methods is essential in order for students to gain a fuller appreciation of the literature underpinning their later academic, or professional careers. Thus, there appears to be a clear and growing need to identify teaching strategies that are maximally effective at removing barriers to learning research methods. This view is echoed by recent calls to reform traditional methods for teaching research methods and statistics, and it finds support from recent research. For example, in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE; Aliaga et al., 2005) college report, published by the American Statistical Association, a number of recommendations are highlighted with regard to the teaching of statistics in higher education. These recommendations include emphasizing the development of statistical literacy and thinking, making use of real data, focusing on conceptual understanding (rather than procedures or formulae), promoting active learning, making use of technology and administering assessment appropriate to evaluating learning in the classroom.

The view that teaching research methods and statistics may require a particular kind of approach is further supported by a recent meta-analysis by Freeman et al. (2014). In their analysis, traditional methods of teaching statistics (e.g., lecturing to classes) was shown to be less effective in terms of student exam performance, and student satisfaction and enjoyment, compared to other subjects of study. The challenge facing teachers of statistics and research methods therefore is to make research methods more applied, relevant and engaging for students, whilst simultaneously improving students' understanding of statistics, their grades, and attendance rates (Hogg, 1991; Lovett and Greenhouse, 2000). In this article, we focus on the possible benefits of implementing two of the recommendations highlighted in the GAISE report. These are: (1) the use of real data, and (2) the use of an active learning methodology. We describe a study that examines the ways in which incorporating these recommendations into the teaching of research methods and statistics may positively affect student outcomes.

When applied to the teaching of research methods, active learning approaches typically involve students carrying out research, rather than merely reading about, or listening to instructors talk about it. Active learning in research methods and statistics classes may include taking part in demonstrations designed to illustrate methodological and statistical concepts, participating in authentic research, and working with data the students have been responsible for collecting. A great deal of work has explored the impact of active learning using 'hands-on' demonstrations of both statistical processes (e.g., Riniolo and Schmidt, 1999; Sciutto, 2000; Christopher and Marek, 2002; Fisher and Richards, 2004) and methodological concepts (e.g., Renner, 2004; Eschman et al., 2005; Madson, 2005). Importantly, the use of active learning methods in research methods and statistics appears to be successful at increasing levels of satisfaction and enjoyment and reducing failure rates (Freeman et al., 2014). Against this backdrop of findings, it might then seem reasonable to assume that the effects of active

learning would further contribute toward positive outcomes, for example on exam performance. However, this is not found to be the case. While students may report higher levels of enjoyment and usefulness of active learning demonstrations, these are not consistently associated with more beneficial learning outcomes (Elliott et al., 2010, though see also Owen and Siakaluk, 2011). Put another way, the subjective evaluation of one's enjoyment of a subject does not bear a direct relationship on the amount of knowledge acquired, or the extent to which one can apply knowledge in a given area (see e.g., Christopher and Marek, 2002; Copeland et al., 2010).

With regard to the use of real datasets in class exercises and assessments, this too has been proposed to hold a number of advantages (Aliaga et al., 2005). The advantages include: increased student interest; the opportunity for students to learn about the relationships between research design, variables, hypotheses, and data collection; the ability for students to use substantive features of the data set (e.g., the combination of variables measured, or the research question being addressed) as a mnemonic device to aid later recall of particular statistical techniques; and the added benefit that using real data can provide opportunities for learning about interesting psychological phenomena, as well as how statistics should be calculated and interpreted (Singer and Willett, 1990). Additionally, a number of studies have showed that when real, class-generated data are used students report higher levels of enjoyment, an enhanced understanding of key concepts, and are likely to endorse the use of real data in future classes (see e.g., Lutsky, 1986; Stedman, 1993; Thompson, 1994; Chapdelaine and Chapman, 1999; Lipsitz, 2000; Ragozzine, 2002; Hamilton and Geraci, 2004; Marek et al., 2004; Morgan, 2009; Neumann et al., 2010, 2013).

Overall, the benefits of using active learning and real data within research methods and statistics classes show much promise. However, to better understand how the implementation of these strategies results in positive outcomes, further empirical investigation is needed. First, we note a lack of research that has simultaneously targeted outcomes of satisfaction, evaluation and knowledge (i.e., performance). Each of these outcomes likely plays an important role in influencing student engagement. In this study we assess students on each of these components. Secondly, we eliminate a potential design confound that may have affected previous research, by ensuring highly similar contexts in both our intervention and our control group. The same instructors were used in both instances. In this way, we may be more confident that any effects we observe are more likely due to our manipulation (i.e., active learning versus control), than to student-instructor interactions.

Motivated by a desire to increase student engagement in our undergraduate statistics and research methods courses, we developed a series of activities for a 1.5-h workshop. In each of these activities, students participated in a computer-based psychological experiment, engaged in class discussions and activities around the methods used in the experiment, and then used data generated by the class to run a range of data handling and statistical procedures. In this paper, we describe an evaluation of the first of these workshop activities in terms of (a) its

subjective appeal to students; and, (b) its pedagogic effectiveness. It was hypothesized that, compared to control participants who were provided with the same content, but delivered using a didactic presentation and canned dataset, students who participated in the activity-based (active learning + real data) workshop would (H1) evaluate the workshop more favorably; (H2) report higher levels of satisfaction with the workshop; (H3) achieve higher scores on a short multiple-choice quiz assessing their knowledge of key learning concepts addressed in the workshop; and (H4) report significantly higher confidence about their ability to demonstrate skills and knowledge acquired and practiced in the workshop.

MATERIALS AND METHODS

Design

A non-equivalent groups (quasi-experimental) design was employed in this study, with intact tutorial classes randomly assigned to the two workshop versions. These workshop versions were equivalent in content, but differed in delivery format. The activity-based version of the workshop began with a computer-based experiment in which the students participated, and contained activities that required students to analyze data collected in class. The canned dataset version of the workshop differed in that it began with a short description of the computer-based experiment (presented by the same instructors as the activity-based workshop), but was otherwise equivalent to the activity-based workshop. As much as possible, the workshops were identical in all other respects. The independent variable in this study was workshop type, of which there were two levels: activity-based and didactic/canned. The four dependent variables were: (1) evaluations, (2) overall satisfaction, (3) knowledge, and (4) confidence.

Participants

Participants were recruited from a participant pool, within which students are required to participate in at least 10 points worth of research during each semester (or complete an alternate written activity). One point was awarded for participating in the current study. A total of 39 participants were obtained for final analysis. Initial comparisons between the activity-based group ($n = 25$; M age $= 22.43$, $SD = 4.95$; 68% female; M final grade $= 61.12$, $SD = 14.54$) and the didactic/canned group ($n = 14$; M age $= 25.93$, SD age $= 12.27$; 78.6% female; M final grade $= 61.42$, $SD = 11.90$) indicated that there were no reliable group differences in age, $t(15.59) = -1.22$, $p = 0.230$, $d = 0.37$, gender distribution, χ^2 (1, $N = 39$) $= 0.50$, $p = 0.482$, $\varphi = 0.11$, or final semester grades, $t(36) = -0.066$, $p = 0.948$, $d = 0.02$.

This research complies with the guidelines for the conduct of research involving human participants, as published by the Australian National Health and Medical Research Council (The National Health, Medical Research Council, the Australian Research Council, and the Australian Vice-Chancellors' Committee [NH&MRC], 2007). Prior to recruitment of any participants, the study was reviewed and approved by the Human Research Ethics Committee at Curtin University.

Consent was indicated by the submission of an online evaluation questionnaire, as described in the participant information immediately preceding it.

Materials and Measures
Workshop

The activity-based version of the workshop commenced with students participating in a short computer based experiment designed to examine the effects of processing depth on recall. Class members were randomized to one of two processing conditions, *imagine* and *rehearse*, then asked to remember a list of 12 words presented on screen at a rate of one word every 2 s. Members of the *imagine* condition were encouraged to engage in deep processing by being instructed to "try to imagine each concept as vividly as possible such that you are able to remember it later." Members of the *rehearse* condition were encouraged to engage in shallow processing by being instructed to "try to rehearse each word silently such that you are able to remember it later." All students then completed multiplication problems for 150 s as a distractor task. Finally, all students were presented with 24 words, 12 of which were 'old' (i.e., appeared on the original list) and 12 of which were 'new'. They were asked to indicate whether each of the 24 words was 'old' or 'new' by pressing a relevant keyboard button.

This task was developed in Java by the second author, as existing commercial software packages were unsuitable for our purposes due to high annual licensing fees (e.g., St James et al., 2005), or an insufficient feature set (e.g., Francis et al., 2008). It was hosted on a private webserver, and accessed by students using a standard web browser (e.g., Firefox). The data generated by each student were saved to a MySQL database accessible to the class tutor from his/her networked workstation. Following their participation, students were provided with a brief written summary of the experiment, and asked to work together to address a series of questions about its key methodological features. These questions prompted students to identify and operationalize independent and dependent variables, write research and null hypotheses, visualize experimental designs using standard notation, and consider the purpose of randomization.

While the students worked on these questions, the tutor downloaded the class data and collated them into an SPSS data file that was subsequently uploaded to a network drive for students to access. After a brief class discussion around the methodology of the experiment, students were directed to open the SPSS data file, and commence work on a series of questions requiring various data handling techniques and statistical analyses to address. Specifically, students were required to identify the appropriate statistical test to compare the two conditions on classification accuracy, and then run, interpret and report (in APA style) an independent samples t-test (including assumption testing, and an effect size). The workshop concluded with a class discussion around the statistical analyses, findings and interpretation.

The didactic/canned version of the workshop was identical to the activity-based version, except it began with a short description of the computer based experiment (presented by the class tutor

with the aid of PowerPoint slides), and required students to analyze a canned data set, rather than class generated data.

Evaluation Questionnaire

The online evaluation questionnaire contained five sections, measuring the four DVs and capturing key demographic data. It is reproduced in full in the Appendix (available as Supplementary Material Data Sheet 1).

Section 1 (evaluations)

Section 1 of the online questionnaire contained 13 items assessing students' evaluations of the workshop. Although there are numerous measures that have been developed to allow students to evaluate units and courses, a review of the literature indicated that there are currently no instruments suitable for evaluating specific activities embedded within a unit or course. Consequently, this measure was developed specifically for the purposes of the current research (although inspired by the single-item measures that are frequently used in evaluations of teaching activities reported elsewhere). Participants responded to each item on a 7-point scale ranging from 1 (*Strongly disagree*) to 7 (*Strongly agree*), and examples of items on this measure include "this workshop was useful" and "this workshop was an effective way of teaching research methods and statistics." Although a small sample size limited our ability to examine the factor structure of this measure (for example, Pett et al. (2003), suggest a minimum of 10–15 cases per item for exploratory factor analysis), Cronbach's alpha was 0.96, indicating that it was internally consistent. Responses to the 13 items were summed to provide an overall index of how favorably students rated the workshop.

Section 2 (satisfaction)

The second section of the online questionnaire was a single item measure of overall satisfaction with the workshop, which respondents answered on a scale ranging from 1 (*Very Dissatisfied*) to 10 (*Very Satisfied*). The correlation between this single item measure and the sum of responses to the 13-item evaluation scale was $r = 0.91$, suggesting that they measured overlapping constructs.

Section 3 (knowledge)

Five multiple-choice questions were used to assess knowledge of the key learning outcomes addressed in the workshop. Each question provided four response options, of which only one was correct, thus total scores on this measure ranged from 0 to 5.

Section 4 (confidence)

This section of the questionnaire asked respondents to indicate on a 4-point scale ranging from 1 (*Not at all confident*) to 4 (*Very confident*) their confidence regarding their ability to apply seven specific skills developed in the workshop, assuming access to their notes and textbook. For example, "run and interpret and independent samples t-test using SPSS." Again, the small sample size limited our ability to examine the factor structure of this measure, although Cronbach's alpha was 0.84, indicating that it was internally consistent. Responses to the items on this measure were summed to provide an overall index of student confidence.

Section 5 (demographics)

The final section of the evaluation questionnaire asked students to specify their age, gender, and the day/time of the workshop they attended. The day/time information was used to assign participants to the levels of the independent variable.

Procedure

Before the start of semester, tutorial classes were block-randomized to the two workshop versions. The workshop was then delivered as part of the normal tutorial schedule. Participants were provided with an information sheet outlining the nature of the current study, and it was stressed that their involvement was (a) entirely voluntary, and (b) anonymous to the unit's teaching staff. At the end of the workshop, students were reminded about the research, and asked to complete the online evaluation questionnaire, which was linked from the unit's Blackboard site, within 48 h of the class finishing. Prior to accessing the online questionnaire, participants were presented with an online version of the information sheet hosted on our school website, as recommended by Allen and Roberts (2010).

RESULTS

Each hypothesis was tested with a Generalized Linear Mixed Model (GLMM), implemented via SPSS GENLINMIXED (version 22), with an alpha level of 0.0125 (to protect against the inflated risk of making Type 1 errors when conducting multiple comparisons on a single data set), and robust parameter estimation. GLMM is preferable to a series of independent samples t-tests or ordinary least squares (OLS) regression analyses, as it can accommodate dependencies arising from nested data structures (in this instance, 39 students nested in seven classes, facilitated by three tutors), non-normal outcome variables, and small, unequal group sizes. In each GLMM, there were two random effects (class and tutor)[1] and one fixed effect (condition) specified. A normal probability distribution was assumed for each outcome variable, and each was linked to the fixed effect with an identity function.

The fixed effects from the four GLMMs are summarized in **Table 1**, where it can be seen that members of the activity-based condition scored significantly higher than members of the didactic/canned condition on the knowledge and confidence measures, but not the evaluation and satisfaction measures. When indexed using Hedges' g, the knowledge and confidence effects could be characterized as 'large' and 'small,' respectively.

---

[1]Note that for five of the eight tests of random effects, the variances were negative, and consequently set at zero during analyses. For iterative procedures (e.g., maximum likelihood estimation), this can occur when the variance attributable to a random effect is relatively small, and the random effect is having a negligible impact on outcome of the analyses. The remaining three random effects were non-significant, with Wald's Z ranging from 0.298 to 0.955 ($p = 0.765$ to 0.340). Despite their non-significance in the current context, the random effects of class and tutor were retained in our analyses, based on the common recommendation that non-independence of observations attributable a study's design ought to be routinely accounted for, regardless of the estimated magnitude of its impact (Murray and Hannan, 1990; Bolker et al., 2009; Thiele and Markussen, 2012; Barr et al., 2013).

TABLE 1 | Summary of differences between the two conditions on the four outcome variables.

Outcome	Estimated condition M (SD)		95% CI	t	p	g
	Activity-based	Didactic/canned				
Evaluations	4.77 (1.22)	4.71 (0.31)	[−0.59,0.72]	0.21	0.836	0.06
Satisfaction	6.68 (2.86)	7.00 (0.95)	[−2.02,1.37]	−0.38	0.703	0.13
Knowledge	3.72 (0.32)	3.43 (0.11)	[0.22,0.36]	8.07	<0.001	1.07
Confidence	2.38 (0.20)	2.33 (0.12)	[0.04,0.08]	5.86	<0.001	0.28

95% CI = 95% confidence interval of the difference between two means. g = Hedges' g for the weighted standardized difference between two means. N = 39 for all outcomes except Satisfaction, where N = 38.

DISCUSSION

We have focused on the implementation of two recommended strategies for teaching research methods and statistics: using real data, and following an active learning approach. Our results showed no reliable differences between groups in their rated evaluation of (H1), or satisfaction with (H2) the workshops. Those participants in the activity-based workshop were statistically no different in their views to those in the didactic/canned workshop. Indeed, it is interesting to note that both groups rated the workshops to be below-average (i.e., below the neutral-point) on the evaluation and satisfaction measures, suggesting that their views regarding the workshops were somewhere between ambivalent and negative. Overall, these findings were not as we predicted. Rather, we expected students in the activity-based workshop to find more satisfaction with their workshop and evaluate their learning experience more favorably. In-line with our predictions, however, was the finding that on the outcome measure of knowledge/performance, the activity-based group did significantly outperform those in the didactic/canned workshop (H3). Thus, while the groups did not differ in their apparent engagement, they nevertheless achieved different levels of knowledge. Also noteworthy, was the finding that the activity-based group were reliably different to the didactic/canned group in their reported levels of confidence to later apply the skills developed in the workshop (H4).

Seemingly, the results of this study sit at odds with the 'causal chain' we described in the introduction. One possible explanation is that for student satisfaction to be positively affected, students need to see the results of their engaged learning first, and perhaps these positive attitudes require time to accumulate. In our study, participants did not have this opportunity. A more interesting possibility is that rather than greater engagement being instrumental in promoting greater levels of satisfaction and enjoyment, which in turn promotes learning, that instead, one's level of satisfaction is in fact rather separate to the process of learning. If so, this would indicate that a combination of teaching strategies is needed to produce positive outcomes and student engagement. Accordingly, our results would be consistent with previous research that suggests exposure to research methods and statistics in an engaging environment can improve students' knowledge without necessarily affecting their attitudes (e.g., Sizemore and Lewandowski, 2009). This latter interpretation offers up a variety of potentially interesting research avenues. Minimally, the results of this study suggest against the tailoring of content in educational curricular, based on the reported levels of satisfaction of students.

Limitations

While the results of the current study raise intriguing questions about the relationship between academic outcomes and self-reported student satisfaction and evaluations, it is important to note a number of possible limitations to the approach we took. The first of these concerns the relatively small, unequal number of participants in the activity-based ($n = 25$) versus canned/didactic ($n = 14$) groups. Clearly, to be more confident in our results, this study requires replication with a larger, more evenly spread sample. A second sampling limitation concerns the randomization of intact groups to conditions. Ideally, we would have randomized individual participants to either the activity-based or didactic/canned workshop, allowing for a true experimental test of each hypothesis. However, this was not possible due to the fact that students self-select into classes based on personal preferences and commitments.

A further possible limitation concerns the analytical approach we chose. Had we opted for another approach, for example independent samples t-tests, no reliable differences would have emerged (ps 0.385–0.839) and the implications of our study would be quite different. However, due to the fact that participants were recruited across a number of tutorial groups ($n = 7$) supervised by a number of instructors ($n = 3$), we deemed the use of GLMM procedures to be most appropriate. This is because GLMM is aptly suited to dealing with hierarchical data, and clustering effects that may have been present within nested groups of tutorials and instructors. GLMM has the further advantage over the t-test in that it may be more robust to dealing with unequal sample sizes (Bolker et al., 2009). Although our analysis showed no such clustering effects, in light of the sampling limitations, GLMM remained most suited to the data.

CONCLUSION

This paper describes the implementation and quasi-experimental evaluation of a relatively short (1.5 h) class activity in which students participated in an authentic computer-based psychological experiment, engaged in class discussion around its methods, and then used class-generated data to run a

range of data handling procedures and statistical tests. Results indicated that students who participated in this activity scored significantly higher than participants in a parallel didactic/canned class on measures of knowledge and confidence, but not on overall evaluations or satisfaction. In contrast to the view that student satisfaction is paramount in achieving positive learning outcomes, the results of the current study suggest that, at least during some points in the learning process, one's level of satisfaction has little effect. This would indicate that a combination of teaching strategies is needed to produce both positive outcomes and student engagement. Future research that employs large-scale, fully randomized experimental designs may have the best chance of revealing these strategies (Wilson-Doenges and Gurung, 2013).

AUTHORS CONTRIBUTIONS

PA conceived and designed the study and analyzed the data. PA and FB co-authored this manuscript. FB programmed the experimental task used as one level of the IV, wrote the documentation and spreadsheets used by the tutors to aggregate the data for class use, and contributed to the overall design of the study.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00279

REFERENCES

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., et al. (2005). *Guidelines for Assessment and Instruction in Statistics Education: College Report*. Alexandria, VA: American Statistical Association.

Allen, P. J., and Roberts, L. D. (2010). The ethics of outsourcing survey research. *Int. J. Technoethics* 1, 35–48. doi: 10.4018/jte.2010070104

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

Bolker, B. J., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., et al. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24, 127–135. doi: 10.1016/j.tree.2008.10.008

Chapdelaine, A., and Chapman, B. L. (1999). Using community-based research projects to teach research methods. *Teach. Psychol.* 26, 101–105. doi: 10.1207/s15328023top2602_4

Christopher, A. N., and Marek, P. (2002). A sweet tasting demonstration of random occurrences. *Teach. Psychol.* 29, 122–125. doi: 10.1207/S15328023TOP2902_09

Conners, F. A., Mccown, S. M., and Roskos-Ewoldsen, B. (1998). Unique challenges in teaching undergraduate statistics. *Teach. Psychol.* 25, 40–42. doi: 10.1207/s15328023top2501_12

Copeland, D. E., Scott, J. R., and Houska, J. A. (2010). Computer-based demonstrations in cognitive psychology: benefits and costs. *Teach. Psychol.* 37, 141–145. doi: 10.1080/00986281003626680

Elliott, L. J., Rice, S., Trafimow, D., Madson, L., and Hipshur, M. F. (2010). Research participation versus classroom lecture: a comparison of student learning. *Teach. Psychol.* 37, 129–131. doi: 10.1080/00986281003626862

Eschman, A., St James, J., Schneider, W., and Zuccolotto, A. (2005). PsychMate: providing psychology majors the tools to do real experiments and learn empirical methods. *Behav. Res. Methods* 37, 301–311. doi: 10.3758/BF03192698

Fisher, L. A., and Richards, D. St. P. (2004). Random walks as motivational material in introductory statistics and probability courses. *Am. Stat. Assoc.* 58, 310–316. doi: 10.1198/000313004X5482

Francis, G., Neath, I., and VanHorn, D. (2008). *CogLab on a CD: Version 2.0*, 4th Edn. Belmont, CA: Wadsworth.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8410–8415. doi: 10.1073/pnas.1319030111

Hamilton, M., and Geraci, L. (2004). Converting an experimental laboratory course from paper and pencil to computer. *Teach. Psychol.* 31, 141–143. doi: 10.1207/s15328023top3102_7

Hogg, R. V. (1991). Statistical education: improvements are badly needed. *Am. Stat.* 45, 342–343. doi: 10.2307/2684473

Lipsitz, A. (2000). Research methods with a smile: a gender difference exercise that teaches methodology. *Teach. Psychol.* 27, 111–113. doi: 10.1207/S15328023TOP2702_07

Lovett, M. C., and Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *Am. Stat.* 54, 196–206. doi: 10.2307/2685590

Lutsky, N. (1986). Undergraduate research experience through the analysis of data sets in psychology courses. *Teach. Psychol.* 13, 119–122. doi: 10.1207/s15328023top1303_4

Madson, L. (2005). Demonstrating the importance of question wording on surveys. *Teach. Psychol.* 32, 40–43. doi: 10.1207/s15328023top3201_9

Marek, P., Christopher, A. N., and Walker, B. J. (2004). Learning by doing: research methods with a theme. *Teach. Psychol.* 31, 128–131. doi: 10.1207/s15328023top3102_6

Morgan, D. L. (2009). Using single-case design and personalized behavior change projects to teach research methods. *Teach. Psychol.* 36, 267–269. doi: 10.1080/00986280903175715

Murray, D. M., and Hannan, P. J. (1990). Planning for the appropriate analysis in school-based drug-use prevention studies. *J. Consult. Clin. Psychol.* 58, 458–468. doi: 10.1037/0022-006X.58.4.458

Murtonen, M. (2005). University students' research orientations: do negative attitudes exist toward quantitative methods? *Scand. J. Educ. Res.* 49, 263–280. doi: 10.1080/00313830500109568

Neumann, D. L., Hood, M., and Neumann, M. M. (2013). Using real-life data when teaching statistics: student perceptions of this strategy in an introductory statistics course. *Stat. Educ. Res. J.* 12, 59–70.

Neumann, D. L., Neumann, M. M., and Hood, M. (2010). The development and evaluation of a survey that makes use of student data to teach statistics. *J. Stat. Educ.* 18, 1–19.

Owen, W. J., and Siakaluk, P. D. (2011). A demonstration of the analysis of variance using physical movement and space. *Teach. Psychol.* 38, 151–154. doi: 10.1177/0098628311411779

Pett, M. A., Lackey, N. R., and Sullivan, J. J. (2003). *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research.* Thousand Oaks, CA: Sage.

Ragozzine, F. (2002). SuperLab LT: evaluation and uses in teaching experimental psychology. *Teach. Psychol.* 29, 251–254. doi: 10.1207/S15328023TOP2903_13

Renner, M. J. (2004). Learning the Rescorla-Wagner model of Pavlovian conditioning: an interactive simulation. *Teach. Psychol.* 31, 146–148. doi: 10.1207/s15328023top3102_9

Riniolo, T. C., and Schmidt, L. A. (1999). Demonstrating the gambler's fallacy in an introductory statistics class. *Teach. Psychol.* 26, 198–200. doi: 10.1207/S15328023TOP260308

Rottinghaus, P. J., Gaffey, A. R., Borgen, F. H., and Ralston, C. A. (2006). Diverse pathways of psychology majors: vocational interests, self-efficacy, and intentions. *Career Dev. Q.* 55, 85–93. doi: 10.1002/j.2161-0045.2006.tb00007.x

Sciutto, M. J. (2000). Demonstration of factors affecting the F ratio. *Teach. Psychol.* 27, 52–53. doi: 10.1207/S15328023TOP2701_12

Singer, J. D., and Willett, J. B. (1990). Improving the teaching of applied statistics: putting the data back into data analysis. *Am. Stat.* 44, 223–230. doi: 10.2307/2685342

Sizemore, O. J., and Lewandowski, G. W. Jr. (2009). Learning might not equal liking: research methods course changes knowledge but not attitudes. *Teach. Psychol.* 36, 90–95. doi: 10.1080/00986280902739727

St James, J. D., Schneider, W., and Eschman, A. (2005). *PsychMate: Experiments for Teaching Psychology (Version 2.0).* Pittsburgh, PA: Psychology Software Tools.

Stedman, M. E. (1993). Statistical pedagogy: employing student-generated data sets in introductory statistics. *Psychol. Rep.* 72, 1036–1038. doi: 10.2466/pr0.1993.72.3.1036

The National Health, Medical Research Council, the Australian Research Council, and the Australian Vice-Chancellors' Committee [NH&MRC]. (2007). *National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015).* Canberra, ACT: Author.

Thiele, J., and Markussen, B. (2012). Potential of GLMM in modelling invasive spread. *CAB Rev.* 7, 1–10. doi: 10.1079/PAVSNNR20127016

Thompson, W. B. (1994). Making data analysis realistic: incorporating research into statistics courses. *Teach. Psychol.* 21, 41–43. doi: 10.1207/s15328023top2101_9

Vittengl, J. R., Bosley, C. Y., Brescia, S. A., Eckardt, E. A., Neidig, J. M., Shelver, K. S., et al. (2004). Why are some undergraduates more (and others less) interested in psychological research? *Teach. Psychol.* 31, 91–97. doi: 10.1207/s15328023top3102_3

Wilson-Doenges, G., and Gurung, R. A. R. (2013). Benchmarks for scholarly investigations of teaching and learning. *Aust. J. Psychol.* 65, 63–70 doi: 10.1111/ajpy.12011

Teaching Research Methods and Statistics in eLearning Environments: Pedagogy, Practical Examples, and Possible Futures

Adam J. Rock*, William L. Coventry, Methuen I. Morgan and Natasha M. Loi

School of Behavioural, Cognitive, and Social Sciences, University of New England, Armidale, NSW, Australia

Generally, academic psychologists are mindful of the fact that, for many students, the study of research methods and statistics is anxiety provoking (Gal et al., 1997). Given the ubiquitous and distributed nature of eLearning systems (Nof et al., 2015), teachers of research methods and statistics need to cultivate an understanding of how to effectively use eLearning tools to inspire psychology students to learn. Consequently, the aim of the present paper is to discuss critically how using eLearning systems might engage psychology students in research methods and statistics. First, we critically appraise definitions of eLearning. Second, we examine numerous important pedagogical principles associated with effectively teaching research methods and statistics using eLearning systems. Subsequently, we provide practical examples of our own eLearning-based class activities designed to engage psychology students to learn statistical concepts such as Factor Analysis and Discriminant Function Analysis. Finally, we discuss general trends in eLearning and possible futures that are pertinent to teachers of research methods and statistics in psychology.

Keywords: eLearning, pedagogy, research methods and statistics, *Second Life*, virtual worlds

INTRODUCTION

Generally, academic psychologists are aware of students' perceptions regarding the "dull, difficult, and distressing" nature of research methods and statistics (Haslam and McGarty, 2014, p. 1). In fact, there is a substantial body of literature devoted to investigating the effect of research methods and statistics on students' anxiety (e.g., Gal et al., 1997). Academic procrastination resulting from statistics anxiety has been linked to numerous variables including the importance of statistics, anxiety associated with interpreting statistical results, anxiety related to exam and classroom contexts, fear of the statistics lecturer, and fear of asking for assistance (Onwuegbuzie, 2004). Importantly, various studies have suggested that effective teaching practices for reducing students' statistics anxiety include a humorous teaching approach, encouragement from the teacher, and the acknowledgment of anxiety coupled with the introduction of coping strategies (see Pan and Tang, 2004).

Cigdem and Topcu (2015) stated that eLearning has extended into most areas of education provision. Given the ubiquitous and distributed nature of eLearning systems (Nof et al., 2015), teachers of research methods and statistics need to be cognizant of how to effectively use eLearning technologies to engage psychology students.

The aim of the present paper is to discuss critically how the use of eLearning systems may facilitate the engagement of psychology students in research methods and statistics. First, we critically appraise definitions of eLearning. Second, we discuss numerous important pedagogical principles associated with effectively teaching research methods and statistics using eLearning systems. Subsequently, we provide practical examples of our own eLearning-based class activities designed to engage psychology students to learn statistical concepts. Finally, we examine general trends in eLearning and possible futures that are pertinent to teachers of research methods and statistics in psychology.

WHAT IS eLearning?

Numerous scholars define eLearning as a variety of online technologies (e.g., *Second Life*) used to facilitate the acquisition of knowledge (e.g., Lorenzi et al., 2004; Asuncion et al., 2007). Others (e.g., Jamison, 2008) moved beyond this rudimentary definition and formulated a dichotomy consisting of *traditional eLearning* (e.g., reading static hypertext pages) and *non-traditional eLearning* [e.g., interactions with avatars in *virtual worlds* (VWs)]. Tripartite models that distinguish between *basic eLearning* (e.g., online pages with assessment), *interactive eLearning* (e.g., the use of multi-media), and *advanced eLearning* (e.g., VWs populated by avatars) arguably superseded this dichotomy (Chapman, 2010). However, definitions of eLearning will evolve as technology evolves. For example, one may envision a historical moment where universities exist solely in cyberspace. In this scenario, the qualifier 'e' in eLearning would become redundant because all learning would be eLearning and, thus, eLearning would be defined as the accumulation of knowledge (i.e., learning; Reber and Reber, 2001). The aforementioned key definitional elements may be synthesized to produce the following definition: eLearning may be defined as the use of online technologies ranging from reading non-interactive contents pages to interacting with avatars in VWs for the purpose of acquiring knowledge and skills.

However, the aforementioned definition is problematic for a number of reasons. First, this definition assumes *a priori* that static hypertext pages constitute the most rudimentary end of the spectrum of eLearning tools whereas VWs and avatars should be located at the most sophisticated end. However, the question is whether *virtual reality* (VR; e.g., the use of immersive head-sets, data-gloves), rather than VWs, constitutes the most technologically sophisticated eLearning tool to date. Second, this definition does not explicitly include *mobile learning* and, thus, the portable aspect of eLearning. Third, given that the term being defined is *eLearning*, it is appropriate that the aforementioned definition is student-centered and, thus, focused on the acquisition of knowledge and skills rather than teaching. However, eLearning is inextricably bound with underlying pedagogical principles and, thus, any comprehensive definition of eLearning should contain an explicit reference to pedagogy. For instance, the social constructivist model was

a key element of Tavangarian et al.'s (2004) definition of eLearning. Thus, it is noteworthy that the aforementioned definition of eLearning is bereft of a reference to pedagogy. Finally, based on an analysis of research articles and a survey of 43 persons, it appears that disparity exists regarding definitions applied to terms such as eLearning (Moore et al., 2011). For example, Moore et al. highlighted that there is disagreement regarding whether definitions of eLearning should be restricted to web-based technological tools (e.g., Nichols, 2003) or include interactive TV and satellite broadcasts (e.g., Ellis, 2004). Thus, from Ellis' perspective, our aforementioned eLearning definition, with its focus on online technologies, is too restrictive. However, Monahan et al. (2008) argued that eLearning once only referred to learning delivered via electronic means. Importantly, with the inception of the internet, the definition of eLearning expanded and now encompasses entire courses delivered online. Thus, perhaps Ellis' position is somewhat archaic. Taking the aforementioned points into consideration, for the purpose of the present paper we will define eLearning as follows: the pedagogically driven use of mobile and non-mobile web-based technologies ranging from hypertext pages to avatar-populated VWs and virtual realities for the purpose of acquiring knowledge and skills.

PEDAGOGICAL PRINCIPLES ASSOCIATED WITH TEACHING RESEARCH METHODS AND STATISTICS WITHIN eLearning SYSTEMS

McLoughlin and Lee (2008a) argued that eLearning pedagogies in tertiary education are often constrained by learning management systems (e.g., Blackboard, Moodle) that simply replicate instructor- and textbook-centered approaches in an online environment. That is, pedagogies need to be developed that allow teachers and learners to actualize the potential of eLearning tools. Unfortunately, some teachers, who are enthusiastic about the notion of eLearning, may use new digital technologies irrespective of whether such technologies are pedagogically effective, or in the complete absence of pedagogical considerations (Beetham and Sharpe, 2007). Thus, the following caution from Hughes (2008, p. 438) is timely: "Technology, without the pedagogy can be a fetishised and empty learning and teaching experience – stylised but without substance or simply electronic information push." Consequently, the aim of this section is to discuss various pedagogical principles, which are pertinent to the effective teaching of research methods and statistics within an eLearning environment.

The Pedagogy 2.0 and Presence Principles

McLoughlin and Lee (2008b, p. 56) stated that, "Pedagogy 2.0 integrates Web 2.0 tools that support knowledge sharing, peer-to-peer networking, and access to a global audience with socio-constructivist learning approaches to facilitate greater learner autonomy, agency, and personalization." A *social-constructivist*

pedagogical approach conceptualizes students as active learners who construct knowledge through: (1) the lenses of their personal experience; and (2) interactions with their teachers and peers (Farkas, 2012). Thus, according to Farkas, the "sage on the stage" model (i.e., the omniscient lecturer as the focal point) is replaced by a learning *community* whereby teachers and learners *co-create* knowledge.

Pedagogy 2.0 is similar, at least in part, to *Presence Pedagogy*, a method of teaching and learning predicated on social constructivist principles (Bronack et al., 2008). More specifically, Presence Pedagogy advocates the following principles: (1) benefiting from the presence of others; (2) encouraging interaction and facilitating community; and (3) sharing resources (Sanders and Melton, 2010). This model is typically applied in a VWs setting, but it may also serve as a guiding philosophy in the context of online discussion forums (Bronack et al., 2008). Such forums allow students to develop online communities and social support networks whereby peers co-create knowledge and share resources. In addition, the forums allow teachers to facilitate students engaged in the social construction of knowledge. For example, in one of our research methods and statistics eLearning systems, an online forum discussion thread emerged whereby students created, and posted, memes to illustrate particular statistical concepts. One series of memes depicted the popular cultural figure Chuck Norris (i.e., an American martial artist and actor) and included the following catch-cries: (1) "Negative correlation: The more Chuck Norris wants to kill you…the less chance you have of living"; and (2) "Perfect Correlation: X = The amount of times Chuck Norris kicks you…Y = Bone fractures you sustain" (Wendy Robertson, personal communication, Thursday 26 March, 2015). Various memes from this discussion thread were incorporated into subsequent lectures. Thus, eLearning systems may facilitate a reciprocal relationship or self-perpetuating feedback-loop whereby the "sage-on-the-stage" (i.e., the lecturer) invokes popular cultural references to illustrate statistical concepts that, in turn, catalyze a network of students to socially construct knowledge (e.g., create memes) that, in turn, further catalyze the lecturer to incorporate the students' memes into subsequent lectures, and so on.

The Learning as Knowledge Creation Principle

Presence pedagogy, with its focus on interaction as a principal method of *co-creating knowledge*, evokes Hong and Sullivan's (2009) principle of *knowledge creation* via collective effort and innovation-oriented approaches. Hong and Sullivan (2009, p. 615) proposed that learning be defined in terms of knowledge creation, a process in which innovation is highlighted as the principal *instructional design goal*. Within this process, individuals are still active participants in their own learning, however, the emphasis is on the "innovative process of inquiry" (Hong and Sullivan, 2009, p. 614) whereby "something new is created and the initial knowledge is either substantially enriched or significantly transformed during the process" (Paavola et al., 2002, p. 24).

Knowledge creation not only further enhances individual knowledge, but "advance[s] community knowledge as a public product" (Hong and Sullivan, 2009, p. 616) as learners work together to develop their learning in the context of a social process that is participatory (McLoughlin and Lee, 2007). Knowledge creation aims to propel beyond a traditionally teacher-focused system in which teachers impart information to passive, receptive students to a system in which students take a more active and constructive role in their own learning. Thus, the emphasis is on a process in which learners actively work to create (or innovate) a path from a problem to a solution (Amabile, 1983; Hong and Sullivan, 2009).

According to Anderson and Dron (2011), social constructivism endorses knowledge creation as a social process. The sociality of humans is emphasized in social constructivism with the recognition that learning is most productive when the environment encourages a multitude of different perspectives in addition to validation, social discussion, and real-world application (Anderson and Dron, 2011). Knowledge creation is, thus, grounded in this constructivist tradition with its focus on "meaningful…activities to support situated learning and knowing" (Hong and Sullivan, 2009, p. 615). The chief point of convergence for this particular principle, however, is the idea of innovative instruction when building knowledge creating communities.

In order for knowledge creation to be actualized as a new pedagogical strategy, instructional design must develop into "a more innovation-oriented approach" (Hong and Sullivan, 2009, p. 614). Thus, utilizing eLearning environments incorporating innovative technologies such as VWs could facilitate the objective of knowledge creation. In collaboration with the teacher, and rather than simply being passive recipients of requisite knowledge (Paavola et al., 2002), students' statistical acumen can be honed in a *knowledge-creating community* (Hong and Sullivan, 2009) in which everyone can work together to increase understanding and feelings of efficacy. As previously noted, research has demonstrated that statistics anxiety is linked with feelings of apprehension, inadequacy, and concerns regarding ability to grasp statistical concepts (e.g., Onwuegbuzie and Wilson, 2003; Onwuegbuzie, 2004). This anxiety has consequences for student performance and relates to students' perceptions regarding their likelihood of passing or failing. In a knowledge-building community, the teacher, together with students who possess a greater statistical aptitude, can scaffold those students who feel less confident in their ability. This advantageous reciprocal relationship immerses students in an environment in which, by working together, students share and reflect upon their existing knowledge and together create new knowledge.

In order to promote a knowledge creating community, a collaborative assessment task could be developed in which students work together to deepen their understanding of the statistical notion "$p < 0.05$." The logic of null hypothesis significance testing is one that many students struggle to grasp early in their statistics education, so this exercise would provide a medium by which they could enhance their comprehension.

Via a wiki delivered through the learning management system, students working in groups of four would each contribute up to 250 words discussing their current understanding of what $p < 0.05$ means to them. They would be encouraged to consider real world analogies in order to actualize this relatively abstract concept as something more concrete and applicable to their everyday experiences. Once all students have contributed their paragraph, as a group, they would work together to assess and discuss each other's work and provide feedback, improving and building upon each other's knowledge. In this way, the integration of newly created knowledge with existing knowledge occurs (Anderson and Dron, 2011). As Green et al. (2010) stated, the use of collaborative assessment has the potential to result in an *adaptive know-how* coupled with an *emergent know-that*, meaning that by working together, students share and reflect upon their own existing knowledge and together create new knowledge.

The Pedagogy of Desire Principle

A *pedagogy of desire* focuses on neglected aspects of teaching and learning (e.g., joy, happiness, transgression) in order to catalyze the desire to teach and learn and, thus, produce teachers and learners who are imaginative, creative agents (Pignatelli, 1999; Zembylas, 2007, p. 340). This principle is particularly pertinent in light of the observation that for many students the prospect of studying research methods and statistics is "boring" or "terrifying" (Gal et al., 1997). Thus, if learners experience boredom or anxiety, then a teacher of statistics might consider promoting a pedagogy of desire that "…produces and seduces imaginations" rather than creating an environment "associated…with repression and coercion" (Zembylas, 2007, p. 332).

We are mindful that previous research demonstrates that humorous teaching strategies may reduce students' statistics anxiety and promote positive affect (e.g., happiness; e.g., Schacht and Stewart, 1990). The following are two examples of this strategy. In a class demonstration devised by one of us the aim is to elucidate the relationship between the reliability (i.e., consistency) and validity (i.e., accuracy) of psychological tests (e.g., an intelligence or IQ test). This demonstration requires a teaching assistant to function as a volunteer. The teacher informs the volunteer that he or she has developed an innovative new method for measuring a person's IQ. The teacher produces a tape measure and measures the circumference of the volunteer's head. On three separate occasions the teacher demonstrates that the circumference is, for example, 24 inches. Thus, the teacher states, "Let us conclude that our volunteer's IQ is 24." Subsequently, the teacher asserts that, "My innovative measure of IQ is reliable because I obtained the same result on three separate occasions. However, my method is not valid because an inch is not a metric that is interchangeable with an intelligence quotient or IQ score. Thus, if a measure is reliable it does not necessarily follow that it is valid."

In another class demonstration devised by one of us, the objective is to explicate an inferential statistical test referred to as a Pearson's product-moment correlation, which measures the strength of the relationship between two variables. To illustrate the concept of a correlation, one of us invokes the character "Barney" from the American situational-comedy "How I Met Your Mother." The episode in which Barney is outlining the relationship between being hot (i.e., aesthetically pleasing) and crazy is described. As a class, we discuss that Barney is arguing that: (1) the correlation is high (i.e., strong); and (2) the direction of the relationship is positive (i.e., as hotness increases so too does craziness). At this point in the proceedings, students often like to venture anecdotes of their own past romantic relationships with hot and crazy individuals.

Importantly, the aforementioned class demonstrations are typically delivered in an eLearning context, using, specifically, *Adobe Connect*, "a web communication system that provides organizations with web communication solutions for training, marketing, and online teaching and learning" (Karabulut and Correia, 2008, p. 483). The teacher hosts the 'meeting' and, importantly, the students do not require software. Instead, the teacher e-mails a link to the students, which allows one to join the session via the internet.

The "Smooth" Space versus "Striated" Space Principle

Deleuze and Guattari (1987, p. 474) asserted that *striated* or *gridded space* denotes space created and perpetuated by the State apparatus, which is formal, structured and hierarchical (Bayne, 2004). Massumi (1987, p. xiii) stated that, "the closed equation of representation, x = x = not y (I = I = not you)" is illustrative of State thought. In contrast, the *smooth*, *rhizomatic space* of *nomad thought* is a "decentered system of points that can connect in any order and without hierarchy" (Murphy and Smith, 2001, p. 1). The term *rhizome* is derived from botany and refers to "a network…that grows horizontally and discontinuously by sending out runners."

Bayne (2004) applied the concepts of the "smooth" and the "striated" to pedagogical cyberspace (e.g., eLearning systems). According to Bayne (2004, p. 302), the "'e-learning system' which, in defining itself as a space of containment, regulation and efficient progression, functions as a strongly striating element within pedagogical web space." More specifically, we note that eLearning systems often exhibit a striated (i.e., hierarchical) *presentation structure*. For example, an eLearning systems homepage is likely to consist of a group of several elements (e.g., general subject information, study schedule and materials, assessment items, forums). Each element leads to other groups of elements. For example, the "general subject information" element may lead to a group of elements (e.g., welcome, contact, how to purchase statistical analysis software, frequently asked questions).

In addition, the content-area of statistics is hierarchical. For example, analysis of variance is an extension of the *t*-test, and multiple regression/correlation is an extension of bivariate regression/correlation (Aron and Aron, 1999). Consequently, week-by-week research methods and statistics study topics featured in eLearning systems will tend to reflect

this hierarchical characteristic (e.g., the *t*-tests study topic is covered before the analysis of variance topic, which is a special extension of the *t*-test). Thus, approaches to teaching research methods and statistics allow one to engage with a striated pedagogical cyberspace in terms of both presentation structure and content.

In contrast, the online discussion forums of eLearning systems provide an opportunity for students to co-create, and traverse, rhizomatic pedagogical cyberspace. For example, as previously stated, in one of our research methods and statistics eLearning systems, students have used online discussion forums to create memes using popular cultural references (e.g., Chuck Norris) with the aim of elucidating statistical concepts. Each popular cultural reference may be conceptualized as a point of a decentered system, which may connect with other points in a multitude of ways without recourse to order or hierarchy (Murphy and Smith, 2001). For instance, in various memes, our students juxtaposed the Teletubbies (i.e., a children's television program), Mr. Spock (i.e., a character from the science fiction television program and movies, "Star Trek"), and Chuck Norris with the aim of co-creating and sharing knowledge with peers.

The "Lines of Flight" Principle

Deleuze and Guattari (1987) developed the notion of *lines of flight* to refer to escape routes from striated space. A line of flight allows a learner, in the context of a relation to one's self, to cultivate a resistance to codes and powers (Deleuze, 1988) and, thus, be able to think otherwise. Essentially, lines of flight may be conceptualized as "…instances of thinking and acting 'outside of the box,' with a greater understanding of what the box is, how it works, and how we can break it open and perhaps transform it for the better" (Lerner, n.d., paragraph 1).

The notion of escape routes from striated space is reminiscent of Heidegger's (1962) concept of *Dasein*, which may be defined as "Being in the world, characterized…in terms of affective relationships with surrounding people and objects" (Blackburn, 1994, p. 94). *Being-in-the-world* equates to inauthentic being on the grounds that our affective relationships to people or objects function to constrain our cognitions, behaviors, and so on. In order to transition from inauthentic to authentic being, one must escape the influence of the "web" of affective relationships by utilizing one's creativity and volition (i.e., "thinking outside the box"; Heidegger, 1962).

An example that one of us devised with the aim of creating a line of flight within an eLearning system is concerned with the ontology of numbers. The teacher pours a carton of milk into a saucer, writes a cat's name (e.g., "Felix") on a slip of paper and, subsequently, places the paper in the saucer. The teacher says to the class: "Felix initially appeared quite dehydrated but now he seems replenished!" Students invariably laugh and the teacher asks what is humorous about this scenario. The students explain that writing a cat's name on a piece of paper does not constitute a real cat. The teacher responds, "Yes!" The teacher suggests that the linguistic term (i.e., word) "cat" is a signifier that is referentially linked to an object (i.e.,

the signified) in the external world with whiskers, fur, a tail and a tendency to "meow." In addition, the teacher asserts that:

> Feeding milk to a linguistic term is an example of confusing the signifier with the signified. It would seem to follow that I have never seen a number and, in fact, do not know what a number is. Why? If I were to write, for example, "8" on the board, then this would constitute a symbol (i.e., the signifier) that is referentially linked to a number (i.e., the signified). However, to assert that "8" is a number is to confuse the signifier with the signified just like I confused the slip of paper with "Felix" written on it with the physical object in the external world.

This demonstration may be delivered via web-conferencing tools (e.g., Adobe Connect) and creates a line of flight by encouraging students to reflect critically on the nature, essence, and existence of numbers and, thus, statistics.

PRACTICAL EXAMPLES USING VIRTUAL WORLDS

As previously stated, traditional eLearning is often reducible to a "network of static hypertext pages" (Brusilovsky, 1999, p. 19), thereby constraining the learner to engage in repetitive read and click functions (Jamison, 2011). What are needed are emerging eLearning tools that facilitate an innovative student-centered experience that is *interactive* and *immersive*. One eLearning tool that allows teachers to be innovative is a VW, which may be defined as "a computer-simulated persistent spatial environment that supports synchronous communication among multiple users who are represented by avatars" (Jung and Kang, 2010, p. 219). VWs include *ActiveWorlds*, *Forterra Systems*, and *Entropia* (Messinger et al., 2009). Currently, in education, the most popular and mature VW platform is *Second Life* (SL; Warburton, 2009).

The innovative potential of VWs provides an opportunity to reshape pedagogical approaches rather than merely replicate traditional teaching methods (Dreher et al., 2009). However, if one were to use SL to simply simulate a PowerPoint presentation in a lecture theater, then the potential for teaching innovation is neglected in favor of "static communication, a single presentation area, and multi-media integrated from Web 2.0 only" (p. 216); see **Figure 1**. Fundamentally, VWs allow the user to virtually experience an object or event rather than simply read text (Chow et al., 2007).

In comparison with the 2-D web, VWs provide numerous innovative ways to facilitate learning (Boulos et al., 2007). For instance, VWs may be used to provide simulated training with the aid of avatars (i.e., an online personal presence) and 'bots' (i.e., an online presence controlled by a machine rather than a human). Examples include role-play simulation in child psychiatry (Vallance et al., 2014), simulated pediatric dentistry (Papadopoulos et al., 2013), virtual patients teaching medical students communication skills (Stevens et al., 2006), and simulated medical emergencies designed to teach CPR to high school students (Youngblood et al., 2007). Numerous studies
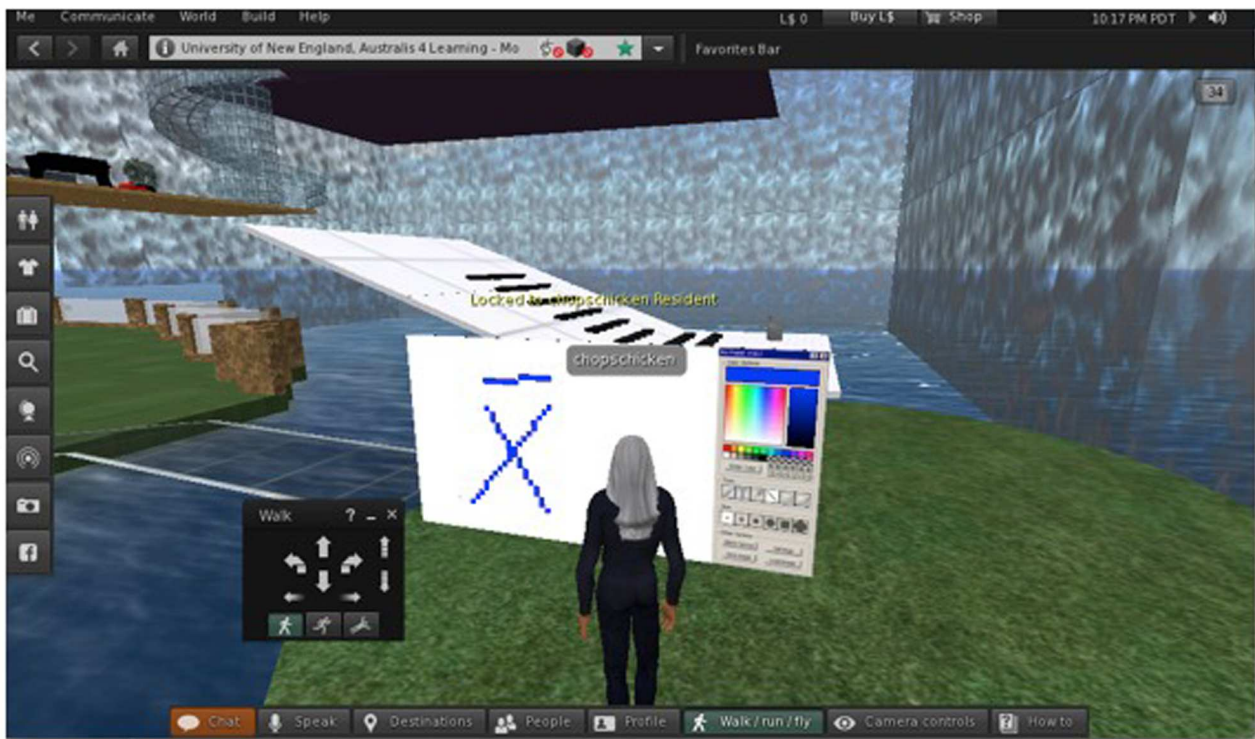
FIGURE 1 | *Second Life* **replicating traditional teaching methods via a virtual interactive whiteboard.**

(e.g., Loftin and Kenney, 1995; Cohen et al., 2013) support the efficacy of simulated training.

In addition, VWs may be used to provide virtual field trips (VFTs), which, via web technologies, can simulate the experience of fieldwork (Arrowsmith et al., 2005). VFTs allow teachers and students to transcend the limitations of time, space, and finances. Garner and Gallo (2005) found no significant differences between a physical field trip group and a VFT group regarding student achievement. The authors concluded that activities such as these can and do promote learning.

Below are two practical examples of statistical methods that can be taught in an engaging and novel way using VWs. We chose VWs because previous research using a VW to engage students in research methods has shown promising results in improving student knowledge and confidence (Baglin et al., 2013). Our two examples focus on statistical tests that are typical of those taught at the 3rd/4th year university level in Psychology, and were chosen because, due to their complexity relative to other statistical methods taught at the same level, they are each better illustrated with a practical example. Providing practical examples in research methods and statistics can be a valuable method to assist students in understanding often abstruse concepts that are difficult to reconcile in the real-world. Research examining the utilization of practical and interesting examples in the teaching of statistics has found that students report a newfound enjoyment for the subject matter as well as seeing an increase in test scores (Burkley and Burkley, 2009). We have delivered these as live class activities for over four years, and the overwhelmingly positive

feedback from students each year affirms they are an effective pedagogical resource.

Practical Example One
Factor Analysis Lesson in a Virtual World such as *Second Life*

Factor analysis is a statistical method used to reduce a large number of variables to a smaller set that best capture the information in the original set. Variables that correlate highly are coalesced into one factor. If multiple factors emerge, then they are structured so as to be largely independent of one another (Cattell, 1952).

The following is a student demonstration designed to provide a rudimentary introduction to the concept of factor analysis in the context of *Second Life*.

(1) The teacher avatar (hereafter "teacher") invites 15 to 20 student avatars (hereafter "students") to stand at the front of the virtual class.

(2) The students are informed they each represent separate variables concerning hair color and, for simplicity, we are interested in the extent to which each variable (or student) correlates with the broader shades of either blonde or dark hair color.

(3) The teacher explains the goal of the demonstration is to reduce the number of variables from 20 to perhaps two or three.

(4) The teacher invites all students with blonde hair to stand together and all students with dark hair to stand together. In so doing, the factor analysis has derived just two factors (i.e., blonde and dark hair), and by using just these two factors the analysis captures a substantial amount of the variation in hair color that was present in the original 20 students. Clearly, having just two factors (or variables) is far more parsimonious than 20.

(5) There will be generally students with brown hair; these students were ignored until now. The teacher then asks whether these students should: (a) be combined with the blonde hair group to create a single blonde-brown group; (b) be combined with the dark hair group to create a single dark-brown group; or (c) form their own group. This allows the teacher to consider what number of factors would be ideal, two or three? The issue is central to factor analysis. For the sake of this demonstration, three factors may be selected.

(6) Typically, there is a student with red or gray hair in the virtual class. The teacher invites these students to walk to the front of the virtual class and join the group to which they belong. However, these students will fit into none of the existing groups. The teacher points out that these students represent an outlier at the variable level. These students are, accordingly, removed from the factor analysis and asked to sit down.

(7) The teacher explains that a factor is a composite of individual variables which all measure the same latent construct. In this example, we have an *amalgam* of various shades of brown that form a single brown hair factor. It is noted that there is a necessary loss of detail in the process of forming the factor. That is, each individual's unique hair color is supplanted by the aggregate brown hair color.

(8) Finally, the teacher explains that these factors are used as predictors in subsequent analyses (e.g., predicting the dependent variable, ethnicity).

Practical Example Two
Discriminant Function Analysis Lesson in a Virtual World such as *Second Life*

Discriminant Function Analysis (DFA) is a statistical method used to predict membership on a categorical (i.e., grouping) dependent variable (DV) from one or more continuous or binary independent variables (IVs). DFA is used when groups are known *a priori*. Thus, the output shows, for each group, the frequencies of the predicted group membership against the actual group membership in order to present intuitively, the prediction accuracy of the analysis (Cohen et al., 2003).

The following is a class activity designed to provide an illustration of DFA in the context of *Second Life*.

(1) The teacher avatar (hereafter "teacher") invites 15–30 student avatars (hereafter "students") to line up in a virtual open space. (One may use between one and three lines depending on the number of students and the size of the virtual space.)

(2) For each line, the teacher nominates a student to be the DFA method "in action."

(3) The teacher invites the nominated students to try and predict, for each student in their line, if each student's father has dark, blonde, or no hair. Fathers' hair type is the DV (i.e., grouping variable).

(4) The teacher explains that the predictions are based on multiple continuous IVs, which include each student's hair color, complexion, and number of hair follicles. Clearly, not all predictions will be correct, which provides a useful illustration of the potential (in) accuracy of the model.

(5) The students are invited to stand in one of three groups that represent whether their father has (or had): (a) dark hair; (b) blonde hair; or (c) is bald. The location of each group is illustrated in **Figure 2**.

(6) Subsequently, the teacher explains that the angle of the first discriminant function, as shown in **Figure 2**, can differentiate between students with: (a) fathers with dark hair from (b) those with either blonde hair or no hair. The variables with a high loading on this function would be student complexion and student hair color.

(7) The teacher explains that the second discriminant function, as shown in **Figure 2**, differentiates (a) the bald group from (b) fathers with hair (dark or blonde). The variable loading high on this function would be the student's number of hair follicles.

(8) The teacher highlights that the two discriminant functions are orthogonal to each other. If there were a third discriminant function it would point directly up in the air.

(9) The teacher reports that the scores on the discriminant functions represent standardized *z* scores, with the mean of zero being in approximately the middle, and high scores being above this and positive, and low scores being below this and negative, as shown in **Figure 2**.

(10) The teacher provides an example of what a standardized value on, for example, the first discriminant function represents. That is, if a student had a high score (i.e., greater than zero) on the first discriminant function they would be in the dark haired group. Importantly, however, if a student had a low score on this first discriminant function they could be in either the blonde or bald group. The teacher explains that it is only by also looking at the student's score on the second discriminant function that we can discern which group they belong to. If a student had a low (i.e., less than zero), rather than high, score on the second discriminant function, and a low score on the first, they would be in the blonde group.

(11) Thus far, we have only attempted to predict group membership. When creating our model, we also need to assess the accuracy of the model by comparing our predictions against the true, rather than predicted, status. We can acquire this information by simply asking each student if his or her father is dark, blonde or bald.

(12) The teacher invites the students who were incorrectly classified to sit down.

(13) Within one group, the teacher explains that the students standing represent the accurate classifications of that group, which can be converted to the percentage correct. This step may be repeated for each of the remaining groups.

(14) Subsequently, the previous step is repeated for all students participating in the demonstration (i.e., the analysis). Thus, the students standing represent the total number of accurate classifications, which can be converted to the total percentage correct.

(15) Finally, the teacher states that DFA models are created with data where the true, rather than predicted, status is known. The goal of DFA is to use the model to generalize beyond the sample in order to predict group status for cases where the true state is unknown. To illustrate this point, the teacher pretends they are an orphan and do not know, or will ever know, their father's identity. However, it is possible to use the DFA model to predict what group the teacher will fall into and, thus, the teacher's father's hair color.

CURRENT TRENDS IN eLearning AND POTENTIAL FUTURES

Martin et al. (2011) analyzed eLearning trends from 2004 to 2014 and identified two themes which we consider pertinent to teaching research methods and statistics:

(1) Mobile devices and the social web are the most important eLearning tools for the near future; and

(2) Games should be deemed an important eLearning tool.

These findings are supported by numerous studies (e.g., Arora et al., 2014; Bhalla, 2014; Yu et al., 2014).

Trend (1) refers to the current shift from eLearning to mobile learning (mLearning), which involves "the use of mobile or

FIGURE 2 | A bird's-eye view of the floor plan for the virtual class demonstration of a Discriminant Function Analysis (DFA). The axes contain standardized values for each discriminant function. The class participants are presented here according to their father's hair type as predicted by the DFA; that is, being bald (at the top), blonde (bottom left), or dark (bottom right).

wireless devices for the purpose of learning while on the move" (Park, 2011, p. 79). An example of a mobile learning technology that pertains to teaching and learning research methods and statistics is StatHand, an application designed to help students cultivate statistical proficiency (About StatHand, 2015; see also Allen et al., 2015). We note that our students have reported using mobile learning devices while engaged in other activities such as horse riding and operating farm machinery (e.g., tractors, harvesters). Such experiences are characterized, in part, by multi-tasking and, thus, divided attention. Fittingly, Lahiri and Moseley (2012, p. 11) cautioned that the use of mobile devices as eLearning tools needs to be underpinned by pedagogical principles and an evidence-base otherwise the use of such tools "might lead to frustration, inequity, shallow learning, and distraction from the main purpose of enhancing learning and making students' competent professionals." Thus, in order to reduce students' statistics anxiety and facilitate student engagement, teachers may wish to consider carefully how to effectively use mobile devices as part of the learning process, which may include the adoption of multiple hitherto unrealized pedagogical strategies (Yu et al., 2014).

Trend (2) refers to the realization by eLearning providers that video game technology can be used to develop fun and immersive simulations (Bhalla, 2014). It is noteworthy that a meta-analysis of game-based learning found that 34 of 65 studies reported statistically significant positive learning effects and only one study reported that computer games were less effective than conventional instruction (Ke, 2009). In addition, a more recent meta-analysis found that, when instructional support was provided, game-based learning enhanced the acquisition of knowledge and skills (Wouters and van Oostendorp, 2013). Trend (2) relates to the goal of facilitating student engagement with statistical concepts. In order to achieve this goal in the context of trend (2), teachers of research methods and statistics need to cultivate an understanding of how video game technologies and principles might be applied in their class. For example, a key principle underpinning the development of video game technology is the facilitation of states of "flow" in the user (i.e., being in the "zone"; Squire, 2003; Cowley et al., 2008; Annetta et al., 2009). If the video game is either too easy or too difficult the user will shift from a flow state to an ordinary waking state characterized by boredom or frustration, respectively (Jamison, personal communication, October 12, 2014). In this regard, we note that in our research methods and statistics computer labs, the proficiencies of students typically fall into three categories: novice, intermediate, and advanced. We have observed that the intermediate students tend to exhibit a flow state. In contrast, the advanced students consider the class too easy and are, thus, bored whereas the novice students regard the class as too difficult and are, thus, anxious and perhaps frustrated. Consequently, the challenge for teachers is to attempt to facilitate flow states in the novice and advanced students. In our own teaching, we have addressed this issue of discrepant learners by delivering separate classes for novice, intermediate, and advanced students. However, we acknowledge the practical issues (e.g., increase in academic workload) associated with such an
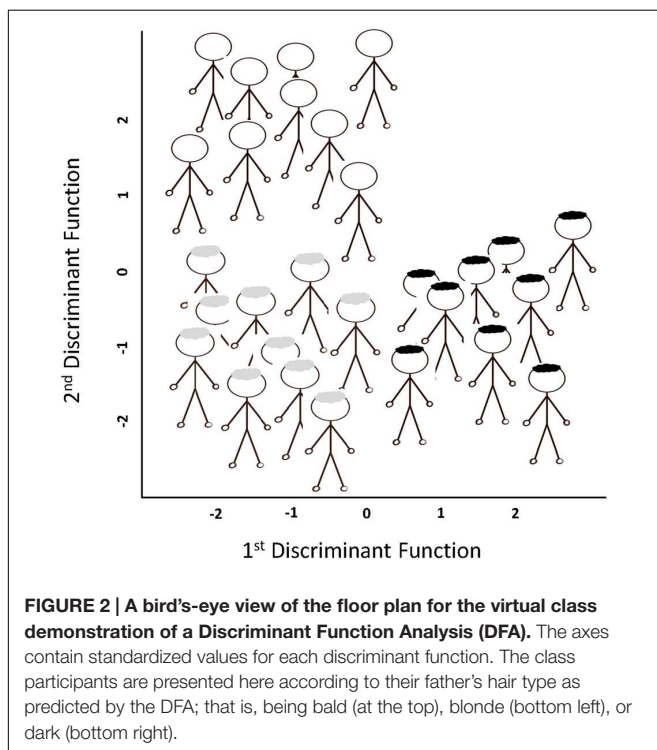
undertaking. Nonetheless, in the context of using eLearning tools to facilitate student engagement with statistics one would be advised to develop tasks designed to optimize the flow states of learners.

CONCLUSION

The objective of the present paper was to examine critically how teachers seeking to engage psychology students in research methods and statistics might use eLearning systems. We demonstrated how various eLearning-related pedagogical principles (i.e., Pedagogy 2.0, Presence Pedagogy, learning as knowledge creation, a pedagogy of desire, striated space versus rhizomatic space, lines of flight) might be applied in the context of teaching research methods and statistics, using examples from our own teaching. Subsequently, we devised two practical examples concerning how Virtual Worlds (e.g., *Second Life*) might be used to deliver class demonstrations concerning two advanced research methods, Factor Analysis and DFA. Finally, we discussed the relevance of mobile learning and video game principles (i.e., the effect of task difficulty on the flow states of the user) to student engagement with research methods and statistics.

In the current era of academic capitalism, which is characterized by the emergence of the entrepreneurial, online university, we note that teachers are constrained to engage in market-like behavior (Slaughter and Leslie, 1997) and provide consumers with *anywhere/anytime learning* (Twining, 2009). Thus, teachers are required to move beyond the notion of the traditional classroom with its face-to-face mode of delivery. In addition, the impending obsolescence of basic eLearning (e.g., students reading static hypertext pages) due to rapid developments in advanced eLearning (e.g., VWs populated by avatars; Chapman, 2010), has resulted in the need for teachers to engage in life-long learning with the aim of maintaining competence in the use of ever-changing eLearning tools and systems. However, we emphasize that the effective use of eLearning tools may be unlikely in the absence of the development of corresponding pedagogies (Hughes, 2008).

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

About StatHand. (2015). Available at: https://www.stathand.net/Home/About

Allen, P., Roberts, L., Baughman, F., van Rooy, D., Rock, A., and Loxton, N. (2015). *StatHand [Computer Software]*. Sydney: Office for Learning and Teaching. Available at: https://www.stathand.net/

Amabile, T. M. (1983). The social psychology of creativity: a componential conceptualization. *J. Pers. Soc. Psychol.* 45, 357–376. doi: 10.1037/0022-3514.45.2.357

Anderson, T., and Dron, J. (2011). Three generations of distance education pedagogy. *Int. Rev. Res. Open Distance Learn.* 12, 80–97.

Annetta, L. A., Minogue, J., Holmes, S. Y., and Cheng, M. T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Comput. Educ.* 53, 74–85. doi: 10.1016/j.compedu.2008.12.020

Aron, A., and Aron, E. N. (1999). *Statistics for Psychology*, 2nd Edn. Upper Saddle River, NJ: Prentice-Hall.

Arora, C., Kaur, C., Gupta, A., and Bhardawaj, A. (2014). A review of recent E-learning trends: implementation and cognitive styles. *Int. J. Inform. Comput. Technol.* 4, 215–220.

Arrowsmith, C., Counihan, A., and McGreevy, D. (2005). Development of a multi-scaled virtual field trip for the teaching and learning of geospatial science. *Int. J. Educ. Dev. Inform. Commun. Technol.* 1, 42–56.

Asuncion, J. V., Fichten, C., and Barile, M. (2007). Which forms of eLearning are accessible to Canadian postsecondary students with disabilities? *Communiqué* 7:36.

Baglin, J., Reece, J., Bulmer, M., and Di Benedetto, M. (2013). "Stimulating the data investigative cycle in less than two hours: using a virtual human population, cloud collaboration and a statistical package to engage students in a quantitative research methods course," in *Proceedings of the Joint IASE/IAOS Satellite Conference*, eds S. Forbes and B. Phillips, Macao.

Bayne, S. (2004). Smoothness and striation in digital learning spaces. *E Learn. Digital Media* 1, 302–316. doi: 10.2304/elea.2004.1.2.6

Beetham, H., and Sharpe, R. (eds) (2007). *Rethinking Pedagogy for a Digital Age. Designing and Delivering e-Learning*. London: Routledge.

Bhalla, S. (2014). E-learning: tools, techniques and trends. *Int. J. Eng. Sci. Invent. Res. Dev.* 1, 82–87.

Blackburn, S. (1994). *Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.

Boulos, M. N. K., Hetherington, L., and Wheeler, S. (2007). Second life: an overview of the potential of 3D virtual worlds in medical and health education. *Health Inform. Lib. J.* 24, 233–245. doi: 10.1111/j.1471-1842.2007. 00733.x

Bronack, S., Sanders, R., Cheney, A., Riedl, R., Tashner, J., and Matzen, N. (2008). Presence pedagogy: teaching and learning in a 3D virtual immersive world. *Int. J. Teach. Learn. Higher Educ.* 20, 59–69.

Brusilovsky, P. (1999). "Adaptive and intelligent technologies for web-based education," in *Kunstliche Intelligenz, Special Issue on Intelligent Systems and Teleteaching*, Vol. 4, eds C. Rollinger and C. Peylo (Cambridge: Academic Press), 19–25.

Burkley, E., and Burkley, M. (2009). Mythbusters: a tool for teaching research methods in psychology. *Teach. Psychol.* 36, 179–184. doi: 10.1080/00986280902739586

Cattell, R. B. (1952). *Factor Analysis, for the Psychologist and Social Scientist*. New York, NY: Harper.

Chapman, B. (2010). *How Long Does it take to Create learning?* Available at: http://www.chapmanalliance.com/howlong/

Chow, A., Andrews, S., and Trueman, R. (2007). "'A 'second life': can this online, virtual reality world be used to increase the overall quality of learning and instruction in graduate distance learning programs?," in *Proceedings of the Association for Educational Communications and Technology International Convention*, ed. M. Simonson (Bloomington, IN: Association for Educational Communications and Technology), 75–83.

Cigdem, H., and Topcu, A. (2015). Predictors of instructors' behavioral intention to use learning management system: a Turkish vocational college example. *Comput. Hum. Behav.* 52, 22–28. doi: 10.1016/j.chb.2015.05.049

Cohen, D., Sevdalis, N., Taylor, D., Kerr, K., Heys, M., Willett, K., et al. (2013). Emergency preparedness in the 21st century: training and preparation modules in virtual environments. *Resuscitation* 84, 78–84. doi: 10.1016/j.resuscitation.2012.05.014

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edn. Mahwah, NJ: Lawrence Erlbaum.

Cowley, B., Charles, D., Black, M., and Hickey, R. (2008). Toward an understanding of flow in video games. *Comput. Entertain.* 6:20. doi: 10.1145/1371216.1371223

Deleuze, G. (1988). *Foucault,* Trans. S. Hand. Minneapolis, MN: University of Minnesota Press.

Deleuze, G., and Guattari, F. (1987). *A Thousand Plateaus: Capitalism and Schizophrenia.* Minneapolis, MN: University of Minnesota Press.

Dreher, C., Reiners, T., Dreher, H., and Dreher, N. (2009). Virtual worlds as a context suited for information systems education: discussion of pedagogical experience and curriculum design with reference to second life. *J. Inform. Syst. Educ.* 20, 211–224.

Ellis, R. (2004). *Down with Boring E-Learning!* (Interview with Dr. Michael W. Allen). Available at: http://www.astd.org/LC/2004/0704_allen.htm

Farkas, M. (2012). Participatory technologies, pedagogy 2.0 and information literacy. *Library Hi Tech* 30, 82–94. doi: 10.1108/07378831211213229

Gal, I., Ginsburg, L., and Schau, C. (1997). "Monitoring attitudes and beliefs in statistics education," in *The Assessment Challenge in Statistics Education*, eds I. Gal and J. B. Garfield (Netherlands: IOS), 37–51.

Garner, L. C., and Gallo, M. A. (2005). Field trips and their effect on student achievement and attitudes: a comparison of physical versus virtual field trips to the Indian River Lagoon. *J. Coll. Sci. Teach.* 34, 14–17.

Green, N. C., Edwards, H., Wolodko, B., Stewart, C., Brooks, M., and Littledyke, R. (2010). Reconceptualising higher education pedagogy in online learning. *Distance Educ.* 31, 257–273. doi: 10.1080/01587919.2010.513951

Haslam, S. A., and McGarty, C. (2014). *Research Methods and Statistics in Psychology.* London: Sage Publisher.

Heidegger, M. (1962). *Being and Time,* Trans. J. Macquarrie and E. Robinson (New York, NY: Harper & Row).

Hong, H.-Y., and Sullivan, F. R. (2009). Towards an idea-centred, principle-based design approach to support learning as knowledge creation. *Educ. Technol. Res. Dev.* 57, 613–627. doi: 10.1007/s11423-009-9122-0

Hughes, J. (2008). "Letting in the Trojan mouse: using an eportfolio system to re-think pedagogy," in *Proceedings of Ascilite, Hello! Where are you in the Landscape of Educational Technology?,* Melbourne. Available at: http://www.ascilite.org.au/conferences/melbourne08/procs/hughes.pdf

Jamison, J. B. (2008). *Educators in a Strange Land: The Experience of Traditional Educators When Immersed into the Virtual Environment of Second Life.* Doctoral dissertation, ProQuest Dissertations and Theses database, UMI No. 3307549, Capella University, Minneapolis, MN.

Jamison, J. B. (2011). *Trance Formational Learning [PowerPoint slides].* Available at: https://prezi.com/zsjq5tdoja6k/tranceformational-learning/

Jung, Y., and Kang, H. (2010). User goals in social virtual worlds: a means-end chain approach. *Comput. Hum. Behav.* 26, 218–225. doi: 10.1016/j.chb.2009.10.002

Karabulut, A., and Correia, A. (2008). "Skype, Elluminate, Adobe Connect, Ivisit: a comparison of web-based video conferencing systems for learning and teaching," in *Proceedings of Society for Information Technology & Teacher Education International Conference 2008*, eds K. McFerrin, R. Weber, R. Carlsen and D. Willis (Chesapeake, VA: Association for the Advancement of Computing in Education), 481–484.

Ke, F. (2009). "A qualitative meta-analysis of computer games as learning tools," in *Handbook of Research on Effective Electronic Gaming in Education*, ed. R. E. Ferdig (New York, NY: Information Science Reference), 1–32.

Lahiri, M., and Moseley, J. L. (2012). Is mobile learning the future of 21st century education? Educational considerations from various perspectives. *Educ. Technol.* 52, 3–13.

Lerner, J. (n.d.). *Lines of Flight.* Available at: http://www.linesofflight.net/linesofflight.htm (accessed April 3, 2007).

Loftin, R. B., and Kenney, P. (1995). Training the Hubble space telescope flight team. *IEEE Comput. Graph. Appl.* 15, 31–37. doi: 10.1109/38.403825

Lorenzi, F., MacKeogh, K., and Fox, S. (2004). Preparing students for learning in an online world: an evaluation of the Student Passport to eLearning (SPEL) model. *Eur. J. Open Distance Learn.* 1. Available at: http://www.eurodl.org/?p=archives&year=2004&halfyear=1&article=108

Martin, S., Diaz, G., Sancristobal, E., Gil, R., Castro, M., and Peire, J. (2011). New technology trends in education: seven years of forecasts and convergence. *Comput. Educ.* 57, 1893–1906. doi: 10.1016/j.compedu.2011.04.003

Massumi, B. (1987). "Translator's foreword: pleasures of philosophy," in *A Thousand Plateaus: Capitalism and Schizophrenia*, eds G. Deleuze and F. Guattari (Minneapolis, MN: University of Minnesota Press), 9–15.

McLoughlin, C., and Lee, M. J. (2008a). "Mapping the digital terrain: new media and social software as catalysts for pedagogical change," in *Hello! Where are you in the Landscape of Educational technology? Proceedings of Ascilite*, Melbourne.

McLoughlin, C., and Lee, M. J. (2008b). The three P's of pedagogy for the networked society: personalization, participation, and productivity. *Int. J. Teach. Learn. Higher Educ.* 20, 10–27.

McLoughlin, C., and Lee, M. J. W. (2007). "Social software and participatory learning: pedagogical choices with technology affordances in the web 2.0 era," in *Proceedings of Ascilite: ICT: Providing Choices for Learners and Learning*, Singapore, 664–675. Available at: http://www.ascilite.org.au/conferences/singapore07/procs/

Messinger, P. R., Stroulia, E., Lyones, K., Bone, M., Nu, R. H., Smirnov, K., et al. (2009). Virtual worlds – Past, present and future: new directions in social computing. *Decis. Support Syst.* 47, 204–228. doi: 10.1016/j.dss.2009.02.014

Monahan, T., McArdle, G., and Bertolotto, M. (2008). Virtual reality for collaborative e-learning. *Comput. Educ.* 50, 1339–1353. doi: 10.1016/j.compedu.2006.12.008

Moore, J. L., Dickson-Deane, C., and Galyen, K. (2011). e-Learning, online learning, and distance learning environments: are they the same? *Int. Higher Educ.* 14, 129–135. doi: 10.1016/j.iheduc.2010.10.001

Murphy, D. W., and Smith, T. S. (2001). What I hear is thinking too: deleuze and Guattari go pop. *Echo* 3, 1–35.

Nichols, M. (2003). A theory of eLearning. *Educ. Technol. Soc.* 6, 1–10.

Nof, S. Y., Ceroni, J., Jeong, W., and Moghaddam, M. (2015). *Revolutionizing Collaboration Through e-Work, e-Business, and e-Service.* Berlin: Springer.

Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assess. Eval. Higher Educ.* 29, 3–19. doi: 10.1080/0260293042000160384

Onwuegbuzie, A. J., and Wilson, V. A. (2003). Statistics anxiety: nature, etiology, antecedents, effects, and treatments – a comprehensive review of the literature. *Teach. Higher Educ.* 8, 195–209. doi: 10.1080/1356251032000052447

Paavola, S., Lipponen, L., and Hakkarainen, K. (2002). "Epistemological foundations for CSCL: a comparison of three models of innovative knowledge communities," in *Computer-Supported Collaborative Learning: Foundations for a CSCL Community: Proceedings of the Computer-Supported Collaborative Learning 2002 Conference*, ed. G. Stahl (Hillsdale, NJ: Lawrence Erlbaum), 24–32.

Pan, W., and Tang, M. (2004). Examining the effectiveness of innovative instructional methods on reducing statistics anxiety for graduate students in the social sciences. *J. Instruct. Psychol.* 31, 149–159.

Papadopoulos, L., Pentzou, A. E., Louloudiadis, K., and Tsiatsos, T. K. (2013). Design and evaluation of a simulation for pediatric dentistry in virtual worlds. *J. Med. Internet Res.* 15, 806–811. doi: 10.2196/jmir.2651

Park, Y. (2011). A pedagogical framework for mobile learning: categorizing educational applications of mobile technologies into four types. *Int. Rev. Res. Open Distance Learn.* 12, 78–102.

Pignatelli, F. (1999). Education and the subject of desire. *Rev. Educ.* 20, 337–352. doi: 10.1080/1071441980200404

Reber, A. S., and Reber, E. (2001). *The Penguin Dictionary of Psychology*, 3rd Edn. London: Penguin.

Sanders, R. L., and Melton, S. J. (2010). The AETZone experience: a qualitative analysis of the use of presence pedagogy in a 3D immersive learning environment. *J. Learn. Teach.* 6, 62–70.

Schacht, S., and Stewart, B. J. (1990). What's funny about statistics? A technique for reducing student anxiety. *Teach. Sociol.* 18, 52–56. doi: 10.2307/1318231

Slaughter, S., and Leslie, L. L. (1997). *Academic Capitalism: Politics, Policies, and the Entrepreneurial University.* London: The John Hopkins University Press.

Squire, K. (2003). Video games in education. *Int. J. Intel. Games Simulat.* 2, 49–62.

Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Raij, A., Harrison, C., et al. (2006). The use of virtual patients to teach medical students history taking and communication skills. *Am. J. Surgery* 191, 806–811. doi: 10.1016/j.amjsurg.2006.03.002

Tavangarian, D., Leypold, M. E., Nölting, K., Röser, M., and Voigt, D. (2004). Is e-learning the solution for individual learning? *Electr. J. E-learn.* 2, 273–280.

Twining, P. (2009). Exploring the educational potential of virtual worlds – Some reflections from the SPP. *Br. J. Educ. Technol.* 40, 496–514. doi: 10.1111/j.1467-8535.2009.00963.x

Vallance, A. K., Hemani, A., Fernandez, V., Livingstone, D., McCusker, K., and Toro-Troconis, M. (2014). Using virtual worlds for role play simulation in child psychiatry: an evaluation study. *Psychiatr. Bull.* 38, 204–2010. doi: 10.1192/pb.bp.113. 044396

Warburton, S. (2009). Second Life in higher education: assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. *Br. J. Educ. Psychol.* 40, 414–426. doi: 10.1111/j.1467-8535.2009. 00952.x

Wouters, P., and van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Comput. Educ.* 60, 412–425. doi: 10.1016/j.compedu.2012.07.018

Youngblood, P., Hedman, L., Creutzfeld, J., Fellander-Tsai, L., Stengard, K., Hansen, K., et al. (2007). Virtual worlds for teaching the new CPR to high school students. *Stud. Health Technol. Inform.* 125, 515–519.

Yu, C., Lee, S. J., and Ewing, C. (2014). "Mobile learning: emerging trends, issues, and challenges in teaching and learning," in *Proceeding of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, ed. T. Bastiaens (Chesapeake, VA: Association for the Advancement of Computing in Education), 2126–2136.

Zembylas, M. (2007). Risks and pleasures: a Deleuzo-Guattarian pedagogy of desire in education. *Br. Educ. Res. J.* 33, 331–347. doi: 10.1080/014119207012 43602

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.

Learning psychological research and statistical concepts using retrieval-based practice

Stephen Wee Hun Lim*, Gavin Jun Peng Ng and Gabriel Qi Hao Wong

Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Singapore, Singapore

Research methods and statistics are an indispensable subject in the undergraduate psychology curriculum, but there are challenges associated with engaging students in it, such as making learning durable. Here we hypothesized that retrieval-based learning promotes long-term retention of statistical knowledge in psychology. Participants either studied the educational material in four consecutive periods, or studied it just once and practiced retrieving the information in the subsequent three periods, and then took a final test through which their learning was assessed. Whereas repeated studying yielded better test performance when the final test was immediately administered, repeated practice yielded better performance when the test was administered a week after. The data suggest that retrieval practice enhanced the learning—produced better long-term retention—of statistical knowledge in psychology than did repeated studying.

Keywords: retrieval-based learning, testing effect, research methods pedagogy, teaching of psychology, experimental education

Introduction

Research methods and statistics are integral to an education in psychology. Ninety-eight percentage of undergraduate psychology programs in North America mandate their students to take at least one methodology class (Stoloff et al., 2009). Psychology graduates who have undergone statistical training acquire critical reasoning skills, distinguishing them from those who have not taken statistics or research methodology classes (Lehman and Nisbett, 1990; Lawson, 1999). Yet, statistics classes can be a source of anxiety (Tremblay et al., 2000) and a dreaded component of the undergraduate psychology curriculum (Conners et al., 1998).

Conners et al. (1998) enumerated four unique challenges for the teaching and learning of undergraduate statistics specifically relating to (a) motivating students, (b) math anxiety (an emotional state of dread toward future math-related activities; see Hembree, 1990), (c) performance extremes and, finally, (d) making learning durable which is of particular interest to the present research. Many educators have noted that students remember very little of what they have previously learned in statistics. One reason is that statistics is akin to a new language, comprising of unique vocabulary and syntax. Lalonde and Gardner (1993) showed that learning statistics is analogous to learning a second language, and argued that it is difficult for students to achieve and maintain fluency with limited exposure. The goal is to discover ways to enhance the learning—increase the retention—of statistical knowledge.

Learning has traditionally been equated to the encoding process through which knowledge is acquired whereas retrieval, often through testing, is viewed as merely a means to judge the extent of prior learning (see, e.g., Karpicke and Roediger, 2008). A fast-growing body of research reveals, however, that retrieval actually aids the retention of previously learned information (e.g.,

Chan and McDermott, 2007; Roediger and Butler, 2011). This phenomenon of improved knowledge recall afforded by retrieval episodes has been referred to as the testing effect (e.g., Carrier and Pashler, 1992), test-enhanced learning (e.g., Roediger and Karpicke, 2006) and, more recently, retrieval-based learning (e.g., Karpicke, 2012).

In the standard retrieval-based learning paradigm, learners either studied educational materials repeatedly, or studied and then practiced retrieving the materials, before taking a final test to assess their learning. In Roediger and Karpicke (2006; Experiment 2), students either studied a prose passage once and underwent three free recall tests about the material, studied the passage three times and took one test, or basically studied the passage four times. They then took a final retention test either 5 min or 1 week later. Traditionally, massed studying produces short-term knowledge retention benefits (see, e.g., Balota et al., 1989). Unsurprisingly, Roediger and Karpicke (2006) found that students who studied the material repeatedly performed better when the retention test was administered immediately. The crucial finding, however, was that students who practiced retrieving performed better when the test was administered 1 week later, implicating the positive effects of retrieval practice on longer-term retention of educationally relevant knowledge (see, also, Gates, 1917).

Lyle and Crawford (2011) implemented the idea of test-enhanced learning in a statistics for psychology course, and found that the student cohort that underwent testing after each lecture eventually obtained higher exam scores than did the cohort which was not tested. While the data imply that testing is advantageous for learning, this advantage is attributable to such reasons as the students in the tested group were simply more motivated to attend lectures—and paid more attention during lectures, since those end-lecture tests were formally graded and students would have taken them seriously. In other words, it is unclear whether the advantage observed was simply due to the fact that the tested cohort basically attended (to) lectures more faithfully than did the untested cohort, rather than due to the prowess of test-enhanced learning *per se*.

The Present Study

Our goal was to illuminate the effects of retrieval-based practice in learning psychological research and statistical concepts under an experimental setting. In line with extant empirical work (e.g., Roediger and Karpicke, 2006; Toppino and Cohen, 2009; Coppens et al., 2011; Kornell et al., 2011) which showed that retrieval-based practice enhances long-term learning, we made two predictions. First, repeated studying—relative to retrieval-based practice—would improve performance when a final test was immediately administered. In contrast, and more important, retrieval-based practice would lead to superior performance in the final delayed test administered after a week.

Materials and Methods

Participants

Sixty-five psychology undergraduates at the National University of Singapore participated for either course credit or a monetary incentive ($10 for an hour of participation). Those who have taken a research methods and statistics course in psychology were excluded from participation. This research was conducted with the appropriate ethics review board approval by the National University of Singapore, and participants have granted their written informed consent.

Materials

A prose passage on the topic of hypothesis testing was developed based on the contents of a textbook chapter by Aron et al. (2009). The passage comprised of concepts in hypothesis testing, central tendency, and decision errors; it contained 361 words, and was decomposable into 26 idea units for scoring purposes.

Design

A 2×2 fully-between design was employed: Participants were randomly assigned to one of two learning conditions: (a) repeated study (SSSS; 36 participants) or (b) retrieval-practice (SRRR; 29 participants). Within each learning condition, about half the participants were assigned to take a final recall test after a 5-min retention interval, whereas the remaining participants took the same recall test after a 1-week retention interval. The dependent variable was proportion of idea units recalled.

Procedure

Participants underwent two sessions. During Phase 1, participants in the repeated study condition studied the passage for four 5-min periods, whereas those in the retrieval practice condition first studied the passage in the first 5-min period and practiced retrieving what they studied in the next three periods, writing down as much material as they could remember from the passage. Participants solved multiplication problems for 2 min in between periods and 5 min at the end of Phase 1. Phase 2 comprised of a 10-min period, during which the final recall test was administered either after 5 min or 1 week later. Participants were asked to recall as much knowledge as they could from the passage administered during Phase 1.
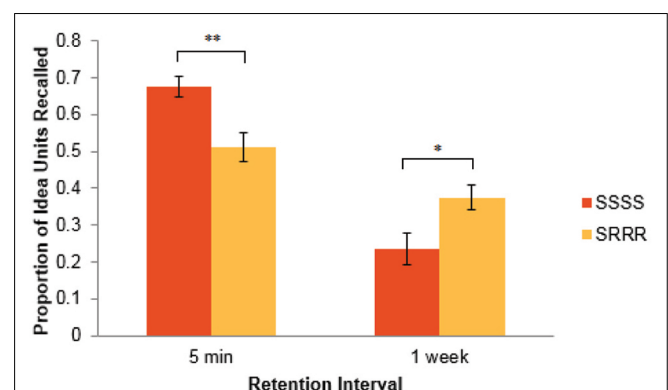

FIGURE 1 | Proportion of idea units recalled across learning condition (*SSSS* versus *SRRR*) and retention interval (5-min versus 1-week). Error bars denote standard errors. *$p < .05$; **$p < .01$.

Results and Discussion

Participants were awarded one point for correctly recalling each of the 26 idea units. The data were then submitted to a 2 × 2 analysis of variance (ANOVA). All assumptions for ANOVA, including independence, normality, and homogeneity of variances, were met. A significant interaction between learning condition and retention interval emerged, $F(1,61) = 17.87$, $p < .001$. Post hoc analyses showed that in the 5-min retention interval condition, repeated study led to a higher proportion of idea units being recalled ($M = 0.675$, SD $= 0.125$) than did retrieval practice ($M = 0.512$, SD $= 0.143$), $t(31) = 3.47$, $p = .002$. In contrast, in the 1-week retention interval condition, retrieval practice led to better recall performance ($M = 0.375$, SD $= 0.130$) than did repeated studying ($M = 0.236$, SD $= 0.173$), $t(30) = 2.58$, $p = .015$. These findings appear summarily in **Figure 1**.

The data supported both of our predictions. While repeated studying, relative to retrieval-based practice, improved recall performance when a final test was immediately administered, retrieval-based practice led to better performance than did repeated studying when the final test was administered after a week. It is worth emphasizing that even though learners who underwent repeated studying read the passage an average of 8.71 times while those who underwent retrieval practice did so only 2.44 times, the latter group was able to recall significantly more idea units after a week has lapsed. Retrieval practice enhances the retention of verbatim knowledge in psychological research and statistical concepts. We have now begun investigating in our Lab whether, and to what extent, retrieval-based learning enhances analogical problem solving — the transfer of previously acquired knowledge or solutions from one context to another — involving psychological research and statistical concepts.

References

Aron, A., Aron, E. N., and Coups, E. J. (2009). *Statistics for Psychology*, 5th Edn. New Jersey: Pearson Education.

Balota, D. A., Duchek, J. M., and Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychol. Aging* 4, 3–9. doi: 10.1037/0882-7974.4.1.3

Carrier, M., and Pashler, H. (1992). The influence of retrieval on retention. *Mem. Cogn.* 20, 633–642. doi: 10.3758/BF03202713

Chan, J. C., and McDermott, K. B. (2007). The testing effect in recognition memory: a dual process account. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 431–437. doi: 10.1037/0278-7393.33.2.431

Conners, F. A., Mccown, S. M., and Roskos-Ewoldsen, B. (1998). Unique challenges in teaching undergraduate statistics. *Teach. Psychol.* 25, 40–42. doi: 10.1207/s15328023top2501_12

Coppens, L. C., Verkoeijen, P. P. J. L., and Rikers, R. M. J. P. (2011). Learning Adinkra symbols: the effect of testing. *J. Cogn. Psychol.* 23, 351–357. doi: 10.1080/20445911.2011.507188

Gates, A. I. (1917). *Recitation as a Factor in Memorizing*. New York: The Science press.

Hembree, R. (1990). The nature, effects, and relief of math anxiety. *J. Res. Math. Educ.* 21, 33–46. doi: 10.2307/749455

Karpicke, J. D. (2012). Retrieval-based learning: active retrieval promotes meaningful learning. *Curr. Dir. Psychol. Sci.* 21, 157–163. doi: 10.1177/0963721412443552

Karpicke, J. D., and Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science* 319, 966–968. doi: 10.1126/science.1152408

Kornell, N., Bjork, R. A., and Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution-based bifurcation model. *J. Mem. Lang.* 65, 85–97. doi: 10.1016/j.jml.2011.04.002

Lalonde, R. N., and Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Can. J. Behav. Sci.* 25, 108–125. doi: 10.1037/h0078792

Lawson, T. J. (1999). Assessing psychological critical thinking as a learning outcome for psychology majors. *Teach. Psychol.* 26, 207–209.

Lehman, D. R., and Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Dev. Psychol.* 26, 952–960. doi: 10.1037/0012-1649.26.6.952

Lyle, K. B., and Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teach. Psychol.* 38, 94–97. doi: 10.1177/0098628311401587

Roediger, H. L., and Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends Cogn. Sci.* 15, 20–27. doi: 10.1016/j.tics.2010.09.003

Roediger, H. L., and Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Stoloff, M., McCarthy, M., Keller, L., Varfolomeeva, V., Lynch, J., Makara, K., et al. (2009). The undergraduate psychology major: an examination of structure and sequence. *Teach. Psychol.* 37, 4–15. doi: 10.1080/00986280903426274

Toppino, T. C., and Cohen, M. S. (2009). The testing effect and the retention interval: questions and answers. *Exp. Psychol.* 56, 252–257. doi: 10.1027/1618-3169.56.4.252

Tremblay, P. F., Gardner, R. C., and Heipel, G. (2000). A model of the relationships among measures of affect, aptitude, and performance in introductory statistics. *Can. J. Behav. Sci.* 32, 40–48. doi: 10.1037/h0087099

Difficult Decisions: A Qualitative Exploration of the Statistical Decision Making Process from the Perspectives of Psychology Students and Academics

Peter J. Allen[1]*, Kate P. Dorozenko[2] and Lynne D. Roberts[1]

[1] *School of Psychology and Speech Pathology, Curtin University, Perth, WA, Australia,* [2] *School of Occupational Therapy and Social Work, Curtin University, Perth, WA, Australia*

Quantitative research methods are essential to the development of professional competence in psychology. They are also an area of weakness for many students. In particular, students are known to struggle with the skill of selecting quantitative analytical strategies appropriate for common research questions, hypotheses and data types. To begin understanding this apparent deficit, we presented nine psychology undergraduates (who had all completed at least one quantitative methods course) with brief research vignettes, and asked them to explicate the process they would follow to identify an appropriate statistical technique for each. Thematic analysis revealed that all participants found this task challenging, and even those who had completed several research methods courses struggled to articulate how they would approach the vignettes on more than a very superficial and intuitive level. While some students recognized that there is a systematic decision making process that can be followed, none could describe it clearly or completely. We then presented the same vignettes to 10 psychology academics with particular expertise in conducting research and/or research methods instruction. Predictably, these "experts" were able to describe a far more systematic, comprehensive, flexible, and nuanced approach to statistical decision making, which begins early in the research process, and pays consideration to multiple contextual factors. They were sensitive to the challenges that students experience when making statistical decisions, which they attributed partially to how research methods and statistics are commonly taught. This sensitivity was reflected in their pedagogic practices. When asked to consider the format and features of an aid that could facilitate the statistical decision making process, both groups expressed a preference for an accessible, comprehensive and reputable resource that follows a basic decision tree logic. For the academics in particular, this aid should function as a teaching tool, which engages the user with each choice-point in the decision making process, rather than simply providing an "answer." Based on these findings, we offer suggestions for tools and strategies that could be deployed in the research methods classroom to facilitate and strengthen students' statistical decision making abilities.

Keywords: statistics, research methods, decision making, selection skills, StatHand, decision tree, graphic organizer, teaching and learning

INTRODUCTION

Quantitative research methods have played a central role in the progress of modern psychology (Benjamin, 2014), and a knowledge of quantitative methods is recognized as essential to the development of psychological literacy (McGovern et al., 2010) and the professional competence of psychology graduates. These points are reflected in the core competencies and graduate attributes specified by accrediting agencies worldwide (e.g., American Psychological Association Board of Educational Affairs Task Force on Psychology Major Competencies, 2013; Australian Psychology Accreditation Council, 2014; British Psychological Society, 2015), and by the prominent position that quantitative methods hold in undergraduate psychology curricula (Perlman and McCann, 1999). This prominence reflects a widely held understanding that an ability to critically evaluate relevant research literature, the vast majority of which is quantitative (Kidd, 2002), is a necessary precursor to evidence-based practice (American Psychological Association Presidential Task Force on Evidence Based Practice, 2006). Engaging students regularly in all aspects of the research process is recognized as fundamental to teaching quantitative methods successfully (Bradstreet, 1996; Stoloff et al., 2015), hence the typical undergraduate psychology degree provides students with multiple opportunities to conduct empirical research, either individually or in collaboration with others (Perlman and McCann, 2005).

Selecting Appropriate Statistics

Despite their prominence and utility, quantitative research methods, and particularly statistics, are known areas of weakness for many psychology students (Garfield and Ben-Zvi, 2007; Murtonen et al., 2008). Students are known to particularly struggle with the development of "selection skills" (Ware and Chastain, 1989, p. 222), or the selection of appropriate statistical tests and procedures for different types of research questions, hypotheses and data types. For example, when Gardner and Hudson (1999) asked students to identify appropriate statistical analyses for a series of brief research vignettes, most found the task extremely difficult, and performed poorly. Even though most had completed at least six research methods and statistics units[1], they managed to identify appropriate statistics for just 25.3% of the scenarios. Gardner and Hudson coded an additional 15.7% of the students' answers as "partially correct." When the researchers questioned the students about how they made their decisions, several explanations for the poor performance emerged. These explanations included students misinterpreting the research scenarios, being unable to actually name known procedures, misidentifying variables' levels of measurement, and answering based on misleading key words and tables of data (which were formatted horizontally rather than vertically, as they would typically appear in a spreadsheet).

If students are required to simply recognize, rather than recall appropriate statistics, their performance is similarly limited. For

example, Ware and Chastain (1989) developed a short multiple-choice selection skill test containing questions pitched at a level they believed a typical student would be able to answer on completion of an introductory statistics unit. However, when they gave the test to students at the conclusion of such a unit, the students answered fewer than 45% of the items correctly. The researchers attributed this poor performance, at least partially, to a curriculum that presented statistical techniques "one at a time" (p. 226), and provided students with few opportunities to practice selection skills. Several other researchers have made similar observations, noting that the typical research methods and statistics unit places far greater emphasis on using known statistical techniques than it does on exploring the circumstances in which they are appropriate (e.g., Bradstreet, 1996; Quilici and Mayer, 1996, 2002; Lovett and Greenhouse, 2000; Yan and Lavigne, 2014). In other words, the difficulties that students experience when placed in situations where they must work out *which technique* to use may be simply attributable to a lack of practice.

When students are provided with opportunities to practice their selection skills, performance increases somewhat (e.g., Ware and Chastain, 1991). For example, when Quilici and Mayer (2002) trained students to focus on the structural features of research scenarios (e.g., the nature of the independent and dependent variables, and the relationship between them), rather than their surface-level characteristics (e.g., the topic of the research), their ability to correctly categorize basic scenarios according to how they would be analyzed improved. The training also improved students' abilities to produce new scenarios with the same structural features as existing ones. However, performance was still far from perfect on both outcome measures. More recently, similar findings were reported by Yan and Lavigne (2014), who also focused their training and categorization tasks on just three basic statistical tests (i.e., independent samples t-test, chi-square test of contingencies, and Pearson's product moment correlation coefficient).

These findings suggest that selection skills are underpinned by a "structural awareness" (Quilici and Mayer, 2002, p. 326), which reflects an ability to disregard the surface features of a research scenario, and instead focus on its structural features and the relations between them. Consider the following section of research vignette four, presented in Appendix A in Supplementary Material:

> *You work at a university library, and have been tasked with finding out which students accrue the largest 'overdue fines'. The head librarian has provided you with a data file that gives you the total amount of fines (in dollars) accrued by each borrower during the previous 12 months, along with a range of additional information (e.g., each borrower's course of study, age, gender, number of items borrowed etc.).*

Identifying an appropriate statistical technique for this scenario requires disregarding its "cover story" or surface-level features, and focusing on identifying its structural features and the relationships between them. In this case, it requires firstly recognizing that the broad intent is prediction (rather than,

---

[1] In the Australian context, a "unit" refers to a single subject, typically taken alongside two or three others over a semester. The term is analogous to "course" in United States higher education parlance.

for example, a comparison between means) and identifying the independent and dependent variables. Here, there are several independent variables of varying types (i.e., dichotomous, nominal, and continuous), and one continuous dependent variable. It secondly involves constructing a generic conceptual model in which the relationships between structural features are represented. In this instance, the intent of the researcher is to use a combination of several independent variables to predict scores on a continuous dependent variable. Finally, it requires integrating the conceptual model with existing knowledge to find possible solutions. For many research scenarios there are a range of statistical techniques that could be used to analyze the data, requiring the researcher to compare possible techniques to determine the most appropriate statistical technique for the particular set of circumstances. While sometimes there may be two or more equally suitable techniques, here the most obvious solution is multiple linear regression, which would provide coefficients useful for addressing the head librarian's question, although additional considerations (e.g., the likely distribution of the dependent variable) may suggest other possibilities. An iterative process may be required between statistical technique selection and testing of assumptions in order to make the final decision.

Without assistance, students find the process described above very challenging. However, "experts" do not. While the point of transition from novice to expert in this specific context is not known, it appears to necessitate a substantial amount of experience. For example, Rabinowitz and Hogan (2008) recruited graduate students enrolled in Masters and PhD courses at a university with "a very well established psychometrics program" (p. 401) to complete a series of triad judgment tasks. In these tasks they were required to identify which of two statistics scenarios "goes best" with a specified target scenario. When faced with the option of selecting a scenario that shared structural but not surface characteristics with the target, or the reverse, even those participants with the greatest amount of experience (i.e., those who had completed between four and eight statistics units previously) did not reliably choose on the basis of structure. Those with the least experience chose based on surface characteristics. Indeed, it was not until the choice was between a scenario that was similar on structural characteristics only and one that was dissimilar on both structure and surface that these "experienced" participants reliably chose based on the structural features of the scenarios. Furthermore, in the Gardner and Hudson (1999) study described earlier, even the most experienced members of their sample (students admitted entry into fourth year, Masters and PhD courses in psychology and education) rarely answered more than 50% of the scenarios they were exposed to correctly.

Beyond the focus on surface and structural components of research scenarios, little is known about how students and experts select statistical tests. The first aim of this research was to develop a rich account of the strategies that psychology students and psychology academics (with expertise in research and/or research methods instruction) use to decide which statistical tests and procedures are appropriate for different research questions, hypotheses and data types.

Decision Making Aids

The preceding section suggests several points. First, even experienced students are not able to autonomously select appropriate statistics in a reliable way. Second, students are often required to make such decisions relatively early in their courses, but are not always explicitly taught how to make them. Third, making such decisions incorrectly can carry substantial negative consequences. At a very pragmatic level, basing a research report on the results of the "wrong" statistical test, will lead to incorrect interpretations and likely poor grades. At a deeper level, it reveals deficits in statistical reasoning or thinking (Bradstreet, 1996; Chance, 2002). Collectively, these points suggest a need for aids or resources that students can rely on to facilitate the statistical decision making process, and perhaps also speed their transition from novice to autonomous expert.

Numerous such aids have been developed, including tip sheets which sort statistical tests according to their defining characteristics (e.g., Twycross and Shields, 2004), and charts which link common research goals to corresponding statistics (e.g., Beitz, 1998). However, the aids which have gained most traction are based around the idea of a "decision tree" or "graphic organizer." Such resources facilitate the decision making process by prompting the user to engage with each structural feature of their research design, as well as the hierarchical and vertical relationships between them (Schau and Mattern, 1997). In the short term, this ensures that the user considers all relevant aspects of the design before deciding on a statistical test, thus increasing the likelihood that a correct decision will ultimately be made. In the longer term, decision trees help users integrate their knowledge of statistical concepts into coherent and organized schemata, which can be quickly and effectively activated when required (Yin, 2012).

Graphic organizers to guide statistical decision making have been used for at least half a century (e.g., Siegel, 1956; Mock, 1972), and are now commonly included in statistics textbooks (e.g., Field, 2013; Tabachnick and Fidell, 2013; Allen et al., 2014). Their inclusion in such books is supported empirically by research on the efficacy graphic organizers generally (e.g., Nesbit and Adesope, 2006) and in the context of statistical decision making specifically. For example, Carlson and colleagues (Carlson et al., 2005; Protsman and Carlson, 2008) demonstrated that graphic organizers could facilitate significantly faster and more accurate (by a multiple of three) statistical decision making, compared to more traditional methods of statistical test selection (e.g., by searching through a familiar textbook). The graphic organizer method was also significantly more popular than the textbook method amongst students.

Regardless of their popularity, traditional statistics decision trees also have a number of limitations. For example, they are often constrained by the requirement that they fit within the pages of a textbook, and when given to students without accompanying resources (e.g., definitions of key terms) they can be of limited use. Koch and Gobell (1999) attempted to overcome this limitation by translating and elaborating a paper-based decision tree for delivery on the world-wide-web. In doing so, they were able to provide students with a range of additional resources, including definitions and information about how to

run and interpret the tests that their online decision tree helped students identify. Like Carlson and colleagues (Carlson et al., 2005; Protsman and Carlson, 2008), Koch and Gobell found that students using their decision tree were better able to identify appropriate statistical tests than students in a comparison condition. Unfortunately, Koch and Gobell's website is no longer active, and many of the online statistical decision trees currently available are of dubious quality or offer little more than could be contained within a traditional paper decision tree.

Aids or resources developed for students to facilitate the statistical decision making process are most likely to be promoted by instructors (experts) and adopted by students if they are developed with expressed needs and preferences of both stakeholder groups in mind. We could locate no research that asked about such needs and preferences regarding statistical decision making aids. Therefore, the second aim of our study was to elicit students' and academics' views on the nature of resources that could facilitate the statistical decision making process.

The Current Study

As noted previously, the two key aims of the current study were to (a) develop a rich account of the strategies that psychology students and psychology academics (with expertise in research and/or research methods instruction) use to decide which statistical tests and procedures are appropriate for different research questions, hypotheses and data types; and (b) elicit students' and academics' views on the nature of resources that could facilitate the statistical decision making process. The study was conducted in two phases. In phase one, undergraduate psychology students were engaged in semi-structured interviews centered on the role and value of statistics, the process of statistical test selection, and the possible characteristics of aids which may facilitate this process. The interpretations from phase one informed the development of phase two. In phase two, psychology academics were engaged in similar interviews, which also queried their perspectives on the challenges students experience when choosing between statistical tests. The findings from both phases will be integrated in the discussion.

This research complies with the guidelines for the conduct of research involving human participants, as published by the Australian National Health and Medical Research Council (National Health and Medical Research Council, Australian Research Council and Australian Vice-Chancellors' Committee, 2007). Prior to recruitment of participants, the study was reviewed and approved by the Human Research Ethics Committee at Curtin University.

PHASE ONE: STUDENTS' DECISION MAKING

Methods
Participants
The phase one participants were nine undergraduate psychology students (five female) with a mean age of 22 years. All had recently completed one or more quantitative research methods and statistics units (median = 3; range = 1–5) and were, on average, in their third year of study. During the interviews, participants were asked to recall their grades for each completed unit, which they did with varying levels of certainty and specificity. When aggregated, these self-reports suggest that the majority of student participants typically achieved "distinction" level grades, with the remainder averaging at the "credit" level[2]. They were recruited via posters placed around university campuses and snowballing.

Materials and Procedure
Data were collected through semi-structured interviews conducted by a research assistant, and guided by a protocol which began by asking participants about the nature of the research methods and statistics units they had taken, and their reflections on those units. They were then directed to a set of brief research vignettes (reproduced in Appendix A in Supplementary Material), prompted to imagine they were the researcher depicted in each, and asked to describe how they would determine appropriate statistics to use. Note that participants were not asked to actually identify a test or procedure (although many did), but rather describe the process or processes they would use to identify one. Following exploration of the vignettes, participants were asked to articulate the reasoning behind the processes they described, and identify processes that others may use in similar situations. Participants were then invited to describe their previous experiences with scenarios like those presented in the vignettes, and prompted to consider the role that an ability to solve such scenarios (or knowledge of an effective process for solving them) plays in a psychology graduate's repertoire of skills. Finally, the interviews concluded by asking participants to describe a tool or resource that they could use to help them approach and solve scenarios like those depicted in the vignettes. The full semi-structured interview protocol is reproduced in Appendix B in Supplementary Material.

Eight interviews were conducted face-to-face, with the final interview conducted via Skype. Each lasted between 30 and 50 min, and was audio recorded for later transcription. Prior to each interview, participants were presented with a participant information sheet, and were given the opportunity to have any questions answered. Face-to-face participants were then asked to sign a consent form, whilst the Skype participant was asked to indicate verbal consent after the consent form had been read aloud by the interviewer. At the conclusion of each interview, and before the recording device was turned off, participants were asked to verbally re-confirm consent, as recommended by Davis et al. (2004).

Data Preparation and Analysis
The audio recordings were transcribed verbatim, and the transcripts were then independently verified for accuracy. The transcripts were imported into NVivo 10, and analyzed following the stages of thematic analysis outlined by Braun and Clarke (2006). Firstly, each transcript was read and re-read, while noting down initial impressions and ideas. Following this initial

---

[2]A "credit" indicates a final mark between 60 and 69%, and a "distinction" ranges from 70 to 79%. For reference a "credit" is typically considered "average" in Australian undergraduate degrees.

familiarization stage, the data were systematically coded in a line-by-line fashion. Codes were then collated into potential themes, which were continually reviewed and refined with reference to the source data and in consultation with team members, colleagues and the research literature. In the final stages of analysis, the themes were defined, and vivid data extracts relating to each were noted for inclusion in this paper.

Findings

Several themes emerged from analysis of the student interview data. Firstly, students overwhelmingly found statistics to be challenging, yet acknowledged their importance for success in a range of different contexts. This is reflected in the theme, *"statistics are challenging, but important."* On the whole, they found identifying appropriate statistical tests for the research vignettes particularly difficult, which resulted in embarrassment for some participants. Many struggled to describe a coherent strategy for approaching the vignettes, however some recognized that approaching them in a coherent and systematic way is possible, and tended to reflect on the utility of flow-charts and decision-trees they had encountered in their studies. These findings are captured by the themes of *"statistical selection falls outside the comfort zone,"* and *"a tenuous grasp on an elusive process."* The students offered a variety of suggestions when prompted to consider the format and features of *"an 'ideal' statistical decision making aid."* Each of these themes is elaborated on in the following sections.

Statistics are Challenging, but Important

Some students indicated that they did not expect to be taught research methods and statistics when they started their psychology degrees (*"it was a bit of a shock initially,"* *"we were so underprepared"*). Others entered the degree with negative expectations about these subjects (*"you hear about statistics before you start psychology and you hear that that's the main reason people drop out"*). They found their early experiences with the subject matter challenging, reporting that there was a lot of *"new"* and *"difficult"* material to learn, and that they sometimes felt *"stressed,"* *"nervous,"* *"confused,"* *"overwhelmed,"* *"overloaded,"* or *"lost."* However, they took some console from knowing that others shared these experiences:

> Everyone's in the same boat …knowing at the very start no one knows what they are doing and everyone feeling a bit lost, it helps you feel like, ah well, I'm not the only one that is having trouble with this.

Many students reported lacking confidence in their abilities (*"I'm just useless at this side of things"*), and that they were not *"math people."* For example, one fourth year student explained, *"I'm a words person not a numbers person, so I was really stressed about doing statistics at uni."* One particular source of anxiety was an exaggerated concern over the consequences of making mistakes:

> Having to figure out what test I was going to use …and still thinking, okay I'm certain, but I'm also a bit unsure. If I pick the wrong test [it will have] a domino effect. Everything else isn't going to work. It …made me feel so nervous.

With experience, the subject matter became more manageable, and students' confidence grew. For example, one third year student remarked that, *"once you've got your foot in the door you can just sort of push through and it's easy."* Having "pushed through the door," research methods and statistics became considerably more enjoyable and rewarding:

> I loved it once I understood it. But just having to go through the stress of trying to understand…getting [tutor] to explain it to me, going over the notes and trying to understand it, getting friends to explain it to me, that was very stressful and that's the part that I just didn't like…But once you actually get a grip on it…I love it!

Despite the challenging nature of the subject matter, students consistently acknowledged the value of research methods and statistics to the development of critical thinking (*"you can question more things, like under what circumstances did they come to that conclusion?"*), to success in their courses, and to competence as future researchers and evidence-based practitioners.

> I'm excited to do honors; to do all the data analysis by myself, and I get to find out things and interpret the numbers. It's like bringing numbers to life, so that's exciting!

> It's important because... psychological research drives all other psychology. It's what forms and guides what every other psychologist will do and practice... or it should do anyway.

Statistical Selection Falls Outside the Comfort Zone

Although we did not ask participants to attempt actually solving the research vignettes, this was the first instinct for many. Most found the task too difficult. They were apologetic and expressed embarrassment at being unable to successfully complete a task they felt they ought to be able to complete:

> I wish I could have done a bit better for you…

> [Interviewer: Do you think that being able to solve problems like these is an important skill for psychology graduates?] Of course, it's a bit embarrassing that I can't do it too well.

However, there was a smaller cohort who jumped straight to a statistic. Occasionally, they did so correctly. Usually though, it was with an unwarranted level of confidence. For example, when presented with a vignette depicting the relationship between two binary variables, a student mid-way through his third year of study answered, *"so it would be a paired samples t-test. Yep that's right. Yep, pretty sure."*

A Tenuous Grasp on an Elusive Process

When prompted to think about the process of selecting a statistic (rather than actually identifying one), students typically struggled. This was the case even for students who had completed several research methods and statistics units:

> [Interviewer: So how would that help you to decide which statistical test to use?] Um see I, see I'm thinking you'd probably want to…I'm sorry. I can't remember, sorry.

The processes they described tended to be haphazard and inefficient, and included looking for (potentially misleading) clues in the wording of the vignettes (*"these scenarios are always worded in certain ways"*), searching through textbooks, lecture notes (*"I would probably just look at …every single test that I've learned about"*), the world-wide-web and previous research addressing similar research questions (*"you've got the journals and things like.…copy their methodology"*). They also reported relying on memory and prior experience or the advice of friends and teachers (*"you could ask your lecturers…'Hey, I'm doing this assignment; what do you reckon I should use?'"*). Some suggested starting by entering their data into a spreadsheet, following a process of elimination, using mnemonic devices or simply guessing:

I kinda try and I guess. I don't know, they're never set in stone, I just kinda think like, 'oh that's probably that one.'

Some students did recognize that a systematic decision making process could be followed: *"you go through checklists in your head."* However, none could identify every factor requiring consideration before an appropriate statistic can be identified. Most also identified irrelevant factors. For example, in the following quote, a fourth year student correctly recognized that she needs to identify the independent and dependent variables (IV and DV), as well as the number of groups being compared. However, she did not consider the measurement levels of the variables (although a nominal IV is implied by her reference to "groups"). Furthermore, she identifies causality as an issue warranting consideration. The appropriateness of causal inference is almost entirely determined by research design, and has very little to do with choice of statistic:

Figure out the variables, the IV, DV I guess. How many groups there are, and what kind of, is it a correlational relationship? …Is it cause and effect?

Those students who recognized a process tended to refer to graphic organizers or decision trees in their statistics textbooks. They reported that such aids facilitated statistical decision making:

The tree! The wonderful tree! It is very simple, easy to use and it pretty much points you right into the analysis that you need to do.

An "Ideal" Statistical Decision Making Aid

Knowing that students find selecting appropriate statistics challenging, we asked those in our sample to explore what might make the process easier. Many turned first to their instructors, who simultaneously helped students master conceptual issues and overcome their hesitation around statistics. When prompted to think about resources they could use independently, technologically based aids were commonly considered:

If you had a website [which] just [asked] how many variables do you have? You know, how many dependent? How many independent?

What are you looking at? What are you comparing to what? And it just tells you this is the test you use.

This idea of a digital decision tree, which focuses the user on a sequence of key decision points before providing a solution was raised often. However, not all students had a preference for digital, with one remaking that she's prefer something in a hard copy format, *"because I can write into it like different things."* Other features of an "ideal" aid included simplicity, accessibility, and multiple levels of depth, as illustrated in the following quotes:

Once you've got the ease-of-use down and you can easily access it, and it tells you exactly what you need to do, I think that's probably all you need really, because once you set it up you can be autonomous and you can self-direct to what you should be doing.

It would be a merge between a super simple tree diagram, but then [a] step-by-step SPSS guide book [and] behind all that a really detailed kind of book …something that comes in three steps: simple, medium and really detailed.

Additionally, students were aware of how the content they access on the world-wide-web is of variable quality, and expressed a preference for content endorsed by recognized "experts," such as *"a psychologist…someone who knows it's going to be useful for other psychologists,"* or *"some Australian government agency."* And finally, an "ideal" aid would contain engaging examples and links to other reputable resources:

Just use like real life examples…like something to do with a person and a situation, instead of saying a group of researchers want to research rats and blah blah.

If there was a way to find more resources…a way to link you with more critical approaches to some statistical tools.

Summary

In the first phase of this study, undergraduate psychology students found our discipline's emphasis on research methods and statistics unexpected, and they approached these subjects with apprehension. They found statistics particularly challenging, but appreciated their importance to success in a range of contexts. Making statistical decisions fell outside the comfort zones of most students, which caused some embarrassment. They had a tenuous grasp on the decision making process, but recognized resources and aids that could guide them through it. When asked to consider the format and features of an "ideal" aid, they expressed a preference for an accessible, comprehensive, and reputable resource that follows a basic decision tree logic.

In the second phase of this study, we turn our attention to the statistical decision making approaches used by psychology academics with particular expertise in conducting research and/or research methods instruction. We also explore their perspectives on the challenges students face when required to choose appropriate statistical tests and procedures, as well as their thoughts about resources that could facilitate this process.

PHASE TWO: ACADEMICS' DECISION MAKING

Methods

Participants

The second phase participants were 10 psychology academics (five female) with appointment levels ranging from lecturer to professor (with a median level of senior lecturer). Six had traditional teaching and research roles, and the remainder were research focused. All were PhD qualified, research active, publishing several papers per year, and supervising research students at the level of honors and above. They predominantly identified as quantitative researchers, although some also used qualitative methods, dependent on the topic of investigation. Half had also coordinated at least one research methods and statistics unit during at least two of the preceding three years. The academic participants were recruited via individual emails, either directly from the first author's professional network, or via colleagues. They were not financially or otherwise compensated for their participation.

Materials, Procedure, Data Preparation, and Analysis

Data were collected through semi-structured interviews conducted by the second author, who did not have a dual role (e.g., as a colleague) with any of the participants. Eight were conducted face-to-face, with the remainder conducted via Skype. As in phase one, all interviews were audio-recorded, following the procedures for obtaining consent described previously. They were guided by protocols (see Appendices C,D in Supplementary Material) that began by querying the functions that statistics play in psychological research and the psychology curriculum. Participants were then directed to the set of research vignettes (presented in Appendix A in Supplementary Material), and asked to describe and explain the process they would use to identify an appropriate statistical test or procedure for each. They were then invited to describe their previous experiences with similar vignettes, and the role that being able to solve them plays in a psychology graduate's repertoire of skills. We then described to participants what we had observed when presenting the vignettes to students in phase one of the study. Specifically, we explained that most of the students struggled to articulate a coherent process, and when they attempted to solve the scenarios they tended to do so incorrectly. We then asked participants why they thought the students found this task so difficult. Finally, participants were asked to describe a tool or resource that students could use to help them approach and solve scenarios like those depicted in the vignettes. Following the interviews, the audio recordings were transcribed, and the transcriptions were analyzed using the techniques described previously.

Findings

Like the students, the academics in the sample also described the importance of statistics, both to their work and the discipline of psychology. They saw "*statistics as a tool*" (amongst several) of research. From their vantage point, the academics also reflected on the nature and value of training in statistics, which they linked primarily to the development of critical thinking and evidence-based practice. This is captured in the theme, "*statistical training underpins competence.*" When prompted to describe the factors that influence their statistical choices, the academics described a complex, nuanced and iterative "*process,*" during which many factors warrant consideration. Some of these factors emerge from the research question and design, whilst others are linked to characteristics of the researcher and broader contextual considerations. These findings are reflected in the theme, "*decision making is a multifaceted process.*" The academic participants recognized that "*students find statistical selection challenging,*" and this knowledge informed their "*pedagogic practices.*" Finally, they described "*an 'ideal' statistical decision making aid*" which shared many of the features identified by the students, but placed a greater emphasis on "the process" rather than "the answer." Each of these findings is elaborated in the sections that follow.

Statistics as a Tool

When asked about the role that statistics play in their work, the academics used terms such as "*central*" and "*vital,*" and suggested that research would be "*pointless*" or "*nothing*" without statistics. However, despite being necessary to quantitative research, being a quantitative researcher requires much more than just knowledge of statistics. To illustrate this point, the "statistics as a tool" metaphor was regularly evoked. For example, "*the way I describe it to students – it's like if you're a tradie or a carpenter, then statistics are your hammer.*" Furthermore, rather than assuming a primary role in the research process, statistics are subservient to the research question and design:

> *The important thing about research, as far as I'm concerned, is not the statistics. That's a tool that you use at the very end in order to answer the question. The important thing in my book is the questions that you're dealing with, that you develop, and the experimental designs that you then use in order to answer your questions.*

In other words, the statistics "fall out" of the design, and the design is a logical consequence of the research question. Or, to quote one of the senior academics in the sample, "*we have a question, we come up with a method of testing it, and we test it and then we move on from there. We get the answer and that the answer is given to us by statistics.*" It is not (or should not be) the reverse:

> *I don't look at it like, 'well I like this statistic, so, I'm gonna design all kinds of studies that I can use this statistic for, or this method for'. I try and look at it the other way around, which is what you're supposed to do.*

Statistical Training Underpins Competence

Participants saw the role that statistics play in psychology curricula as multifaceted, and that a rigorous background in quantitative methods can distinguish the psychology graduate from graduates of other disciplines, ("*that's what makes psychologists or psychology graduates cool and different*"). While noting that statistical literacy was a necessary precondition for conducting research, they saw the primary purpose of statistical

training as tied to the competent consumption (and evaluation) of research literature and the development of critical thinking skills:

> I do think it's a very central skill that they should be able to come out and go, 'Okay. Well, I can read this paper and think they've done the appropriate analysis,' and not have to rely on conclusions the authors have drawn…You're sort of critically consuming information rather than just taking what you're told.

Participants also saw training in research methods and statistics as providing a general framework for applied problem solving: "I think that approaching complex social problems in general requires you to have an understanding of multivariate and quantitative statistics. So it makes you a more informed citizen." Furthermore, the ability to understand and evaluate research literature and solve problems were widely regarded as necessary pre-requisites for evidence based practice: "We base our profession on the scientist-practitioner model, so the evidence base is very important and statistics are really the – what we use to establish that evidence base." However, this sentiment was not universal, with one participant commenting that, "I'm not really aware of any data which suggests that their statistical expertize is associated with better performance as a clinician…Not everyone needs as much [training in statistics and research methods]."

Despite generally recognizing their importance, some participants noted that we do not do a good job of communicating this importance to students, which may be linked to students often only appreciating the relevance of statistics and research methods in hindsight:

> I don't think the reason we include them [statistics] in psych is ever made very clear to students

> The feedback I get from students is often delayed…They come back a year later and say, 'thank you, I really enjoyed that. Now I understand it.' But it's a shame. I wish they would have had that eureka moment a bit earlier …

Decision Making is a Multifaceted Process

When prompted to explicate the factors influencing analytic choices, participants described a complex, nuanced and iterative "process," during which many issues warrant consideration:

> Often there are a number of different ways to answer a question and which one's appropriate depends on the current state of the literature, obviously the data that you've collected, what it is you want to get out of it, where it's going to be published…

This process begins with "the question" and design, followed by the nature of the variables in the study. In fact, the prevailing attitude was that, without a clear research question and intent in mind, any discussion of statistics was premature. For example, when asked about how he would respond to a student who had research ideas, but was uncertain about the appropriate statistics, one participant stated, "I would tell them that they shouldn't worry about stats; they should worry about the questions that they have, how they can operationalize the question, put it into a research

design that will give them an answer, and then we'll worry about the stats later." However, while "jumping" into statistics too soon was regarded as poor practice, so was leaving the development of an analytic plan too long. Doing so can prove costly, as illustrated in the reflections of one senior research focused academic:

> For one of the studies for my PhD I collected a load of data and then realized it actually wasn't analyzable in SPSS …And that's where I started realizing the importance of knowing what you're doing before you start, and not collecting data and then saying, 'well, how will I analyze this?'

When developing an analytic plan, participants most commonly looked to aspects of the study. However, personal characteristics and contextual factors can also play a role in the decision making process.

Characteristics of the Study

Having a clear understanding of the purpose and design of the study as well as the number and nature of variables were recognized as essential to being able to select an appropriate statistic. For example, when presented with the second scenario in Appendix A in Supplementary Material, an experienced research methods instructor explained:

> I see a between groups three level IV. And then I see a between groups two level IV. So I'm thinking a two by three factorial design. And I'm seeing this repeated measures …So at this point I can see there's a choice between - like the way it's written implies that the dependent measure is an average over five trials. So that's a 2×3 between groups design. Of course, you could look at it as a three way mixed ANOVA with 'trial' as a third factor, which allows you to look at trajectories of learning. So I'm thinking if I'm writing for a journal, a learning journal, I'm pretty sure that it would be a three way mixed design. As it's presented here thought it looks like a two by three between groups design.

Participants also noted that consideration should be given to alternative options in the event that analytic plans require modification due to, for example, violated assumptions. The importance of considering Type 1 and Type 2 error rates, statistical power, and the directionality of hypotheses during the decision making process were also discussed. Notably, participants actively considered viable alternatives, and weighed up the benefits and challenges associated with different decisions. This was particularly evident when discussing the mentoring of junior researchers:

> Usually I will try and elicit their ideas first, and then pose some questions if I think there are other options, and ask whether they'd considered them. And if not, why not. Or if they had considered them, but decided on an alternative method, discuss why that is.

There was also a degree of tension between what could be considered "ideal," and what is realistic or possible. As explained by one of the instructors, "there's quite a few different ways to actually do things, of varying levels of effectiveness, and depending on the resources that you have."

Personal Factors

Participants expressed an element of personal preference when considering appropriate analytical strategies ("*I'm not a fan of mixed ANOVAs. I much prefer to go through with repeated measures ANOVAs…*"), although it was recognized that such an approach does not reflect "*best practice.*" There was also some tension between a desire to prove competence and an appreciation that the "best" technique is not necessarily the most complex:

> *There is something nice about really complex designs and really complex analyses that tend to stun people into thinking, 'you know what you're talking about!'*

> *I tend to err on the side of you use the technique that's appropriate, not the fanciest one. So there's something to be said for if a t-test answers your question, use a t-test. Like there's no need to get all fancy just for the sake of it.*

Contextual Factors

It was observed by academic participants that research is not conducted in a vacuum, and that there are factors outside the researcher's immediate control which influence the statistical decisions they make. The first of these is the intended audience: "*What people need to realize is that the choice of analysis is on par with choice of audience… [and] sometimes you have to do different analyses for different audiences.*" As reviewers and journal editors are frequently gatekeepers between researchers and their broader audiences, their opinions were given particular weight: "*Then you get a reviewer who has their own preference on the type of statistics they would like to be used, so you have to revise it.*" At times, these opinions were seen as useful, and helped shape future decision making. At other times, they could be an impediment to progress:

> *I was always taught that if you're testing mediation, you should use Baron and Kenny's model which is now, indeed, 20 years out-of-date, and there are whole books on much better ways of doing it. And the only way I came across that was when I submitted a paper with mediation and one of the reviewers said, 'yeah, this is okay, but there's much more sophisticated and better ways of testing that'. It put me into touch with a whole literature which I now – anytime I'm testing mediation, we use those.*

> *And what I have experienced this last year, actually, is that I did use different statistical methods working with [a statistical consultant]… And because they were different, they were met with – reviewers didn't like it. They didn't like things that they didn't know. So you'd have to explain it, and they thought that you were trying to trip them up or trick them to get something.*

Participants also made regular reference to how shifting discipline practices (and what is considered "best practice") can influence decision making. For example, one participant described how she used simple regression techniques in her PhD. Yet, if she was examining a current PhD in which the same techniques were used, she would say "*no way, go back and do something much, much better.*" Furthermore, although best practice guides decision making, what defines best practice is often quite opaque:

> *There is uncertainty … because there's no black and white. It's not really that kind of field. So you might find one article that said, 'breaking the assumption is okay under these circumstances. You can get away with it.' And in other circumstances you can't. So you often get contradictory messages.*

The preceding quote indicates that there may be a range of "best practices," and what is ultimately acceptable depends both on the technique applied, as well as its justification:

> *With my graduate students, a lot of what I'm teaching is 'yes there are some fundamentals, but once you get beyond that it's about being able to determine the appropriate technique for your question and your data and then be able to justify that decision knowing that you'll send it out for review and people will disagree with you'.*

Finally, beyond an aspiration toward best practice, participants also indicated a desire to avoid (or be seen to avoid) poor practice. The poor statistical practices most commonly cited centered on "fishing" for effects and their subsequent misrepresentation in published work:

> *If you're just doing post hoc analysis, but pretending that it was a priori, then you get – I've seen it at conferences; students claiming they did a mediated moderation on one thing and then moderated mediation on the other. And you kind of go, 'there's no way that was a priori. You did not go into the research with that plan!' If you do enough statistical tests and you don't report them, and you don't do Bonferroni corrections, then you run the risk that something is going to be significant, just because.*

Students Find Statistical Selection Challenging

Aside from a small cohort of particularly capable students, it was widely recognized by the academic participants that many students find research methods and statistics challenging sections of a psychology degree. When we described the outcomes of presenting the research vignettes to the student sample, and asked academic participants why they thought the majority of students struggled with them, a range of possibilities were suggested. Some of these appeared to be attitudes or dispositions that students brought to the degree or developed over time, whereas others reflected characteristics of the teaching methods and materials commonly used in undergraduate psychology courses.

Student Characteristics

Participants perceived that the reality of a psychology degree is often inconsistent with students' expectations on entering the course. This could be because psychology "*doesn't sound like a course that requires a lot of statistics.*" They also noted that many students approach statistics with anxiety, lack confidence in their statistical abilities, are disinterested in research methods and statistics, or do not see their relevance to their future professional lives:

> *Students are scared of statistics. And therefore they get a bit of a mental block, I think, and convince themselves they don't know how to answer the question.*

> *It's perceived as another class they don't like, that they don't perceive is relevant, that they don't understand – It's like math at school,*

'when am I ever going to use this?' Because students coming in are all gong to be clinical psychologists and we know clinical psychologists never use numbers <laughs>!

Course Characteristics

Academic participants highlighted both implicit and explicit characteristics of the research methods and statistics curriculum which may hinder, rather than support students' skill development. For example, one participant described the discipline's tendency to "*fetishize*" statistics, and how this value is communicated to students:

"There's an element of elitism. If we make it seem really hard and difficult to get into and make it really opaque, we're shoring up this idea that stats is for the hard men and the real - we can sort the men from the boys amongst the students and also amongst everyone else of us too."

Others spoke of teaching approaches which tend to compartmentalize content, which is stripped of context when presented to students:

It was very much pigeon-holed. So it was very much this week we're talking about ANOVA; this week, we're talking about regression; this week, we're talking about something else. So there really wasn't that opportunity to make a decision about which one is which. It was just, 'this is what you're doing'.

Overwhelmingly though, participants ascribed the difficulties students have with statistical decision making to teaching methods which don't engage students in regular decision making opportunities from early in the course ("*there just isn't enough exposure to that sequence of thought and planning*"), and don't regularly reinforce the relevance of statistics. It was considered that both these aims could be achieved by engaging students in the full research "process." To participants, this process begins with a substantive research question, works through key issues tied to design and analysis, and concludes with clear implications or, to quote one instructor, an answer to the question, "*what does this shit actually mean?*"

Showing that it's not necessarily about numbers but about answering questions might help with some of the – and putting it into that context, and putting it into the context of a research problem and not a math problem - I think, it can help as well.

Answering questions of substantive interest was seen as vital. Furthermore, failing to achieve this aim may promote disinterest, disengagement, and apathy.

…as soon as it's a question that you wanna know the answer to, it's like …it suddenly becomes relevant and important.

Pedagogic Practices

Recognizing that statistical decision making is an area that students find challenging, participants employed a number of techniques to encourage and support their efforts. This tended to occur in the context of either small-group/individual research

supervision sessions or lab group meetings. Firstly, questioning was used to guide students "*through the process.*"

I use a lot of questioning and I'm just thinking about one student that I spoke to just last week who put point blank to me, she said, 'oh, we'll be using [multiple] regression to answer this question,' and I immediately sort of flicked it back on her and said, 'but how are you measuring your DV?' – which was dichotomous. So in asking that question, she was able to go, 'oh hang on a minute…that data is not appropriate for what I just said'.

The process involves considering design and statistical issues concurrently, and in the context of the research question or objective:

I ask them to draw out the design of an experiment, say, and they might suggest some stats at the end. And then, I ask them how that addresses the question or questions [they] want to get to.

It also involves consideration and evaluation of different options before making decisions, and collaboration and consultation is encouraged:

…try and present the different options…what are the pros and cons of each in this case, and then weigh those and come to a decision. I think you kind of need to let them go through the process.

An "Ideal" Statistical Decision Making Aid

Academic participants suggested characteristics for a tool or resource that students could make use of to independently identify appropriate statistics for various circumstances. First, the resource should be accessible (in terms of ease and cost of availability), and step users through a sequence of questions or decisions which must be addressed to arrive at an actionable outcome. Terms like "*flow-chart*" and "*decision-tree*" were used commonly.

It is a question and answer flow-chart kind of situation. Is it relationships or differences? …how many variables; categorical or continuous? The answers to each of those questions would lead you to the correct [statistical analysis].

It seems like if there was some sort of decision tree …It would make sense to have some sort of app or something …easily accessible online or on your phone or whatever, where you can plug in and go through a step-by-step process.

If questions or decision points are presented sequentially, the user is forced to engage with each step in "*the process*" and can thus be "*train[ed] …to ask the important questions.*" The longer term objective of such a resource should not be reliance, but rather a transition toward greater autonomy and flexibility:

[After using the resource for a period of time, the user should ideally be able to] turn it off or turn the book over and then you give them another problem and see, well can they now - are they now able to - even if they can't get to the right answer, are they now trying to figure out? 'Well, what am I trying to do? How many groups and

what am I - what's my IV, what's my DV, do I have more than one IV, what's the level of measurement?'

Participants also noted that understanding key terms (or having the ability to quickly look them up) is essential to being able to use such a resource effectively ("*you need to know what a covariate is, what the IVs and the DVs, what this actually means*"). Finally, they acknowledged that, realistically, such a resource is never going to capture all the nuances in statistical decision making, but may be useful within the broader discussion:

If you try to reduce it to a few basic principles then you're missing critical questions, like 'what is the hypothesis' and 'what is the audience'? It's really much better if it's a consultative process with an advisor and/or with other [students]. I don't think people should work independently necessarily. I think that there's a lot of virtue in consulting with people in the design phase of the project.

Summary

In this study's second phase, the academics saw statistics as one of several tools available to the researcher; a tool that is vital to the conduct of most research, but subservient to the research question and design. They acknowledged the role that statistics training plays in the development of research skills, but saw its primary role as nurturing the development of critical thinking and evidenced-based practice. The academics described choosing an appropriate statistic as a complex, nuanced, and iterative process, during which consideration should be paid to multiple contextual factors in addition to the characteristics of the study. They were sensitive to the challenges that many students experience when making statistical decisions, which they attributed partially to how research methods and statistics are commonly taught. This sensitivity was reflected in their pedagogic practices. The "ideal" statistical decision making aid the academics described shared many of the features identified by the student participants, although greater emphasis was placed on "the process" than "the answer."

DISCUSSION

The first aim of this research was to explore the strategies that psychology students and academics use to select statistical tests. We probed these strategies in semi-structured interviews, in which participants were encouraged to discuss how they would approach each of a series of short research vignettes. Our findings indicate a number of key differences between how these two groups approach statistical decision making.

For the students in our sample, being required to make such decisions pushed them outside their comfort zones, resulting in either apologetic discomfort, or instinctual selections that were frequently incorrect. This finding is not surprising given the body of literature demonstrating that most students find statistics generally (Garfield and Ben-Zvi, 2007; Murtonen et al., 2008), and statistical decision making specifically (Ware and Chastain, 1991; Gardner and Hudson, 1999) to be difficult. Their ability to even describe the process of selecting a test was limited, and relied heavily on the use of strategies unlikely to produce optimal

outcomes. These included searching through textbooks, lecture notes, and the world-wide-web, relying on memory and prior experience, turning to the advice of friends or teachers, and looking for clues in the wording or structure of the vignettes. A number of these strategies were also suggested or displayed by the students in Gardner and Hudson's (1999) research, who were particularly prone to misinterpreting research questions, and being mislead by key words and data presentations formats. Like those in Gardner and Hudson's research, the students in our sample were reasonably far into their degrees and were, on average, in their third year of study.

There were a minority of students who recognized that a systematic decision making process could be used to approach and "solve" the research vignettes. However, none were able to identify all the factors in the vignettes that would require consideration before appropriate statistics could be identified. Furthermore, these students had a tendency to also identify features of the vignettes which were irrelevant to the task at hand. Again, these findings are broadly consistent with Gardner and Hudson (1999), whose students often failed to take the nature of data (e.g., nominal, ordinal etc.) into consideration when making statistical decisions.

By way of contrast, the psychology academics described selecting appropriate statistics as a complex, nuanced and iterative process, embedded within the broader process of conducting research. They demonstrated how during statistical decision making, consideration ought to be paid to multiple contextual factors (e.g., the intended audience, prevailing discipline trends and practices etc.), in addition to the intent and design of the study itself. These experts were able to suggest appropriate statistical analyses for each vignette with ease, but were often reluctant to do so without understanding the purpose of the research, or having an opportunity to explore alternative possibilities. This behavior is suggestive of "structural awareness," which is an ability to see past the surface features of a problem, and focus on its structural characteristics and the relations between them (Quilici and Mayer, 2002)[3] . It is a characteristic common to "expert" problem solvers across a wide range of specialized domains (Rabinowitz and Hogan, 2008).

Previous research suggests that structural awareness tends to develop naturally with experience (Rabinowitz and Hogan, 2008). In the Australian context, opportunities to engage in statistical decision making are limited prior to fourth year when, under individual supervision, psychology students embark on their first major research project. During this intensive research internship, expert supervisors model the statistical decision making process, and use a range of techniques to promote its

---

[3]Despite this structural awareness, the findings suggest that some psychology research academics have preferred techniques, will at times select techniques based on what they can "sell" rather than current best practice, and are reluctant to be early adopters of new techniques. This "resistance" by substantive psychological researchers to changing statistical techniques and employing new advanced statistical techniques has previously been recognized in the research literature (Sharpe, 2013). It has been attributed to a combination of a lack of awareness of new statistical developments, inadequate statistical education, the failure of journal editors to act as catalysts for change, the pressure to "publish or perish," and fear of deviating from normative statistical practices (Sharpe, 2013).

development in students. Students in earlier years are largely reliant on lectures, laboratories, and tutorials to develop their research skills, and alternative methods of teaching statistical test selection, which are not reliant on individual supervision, are required for these years.

Our recommendation is to provide students with regular opportunities to engage in the statistical decision making process in the context of class research projects. It is widely recognized that scaffolded immersion in all aspects of the research process, from participation and/or data collection, through the development and testing of hypotheses, to the interpretation and reporting of findings, is a particularly effective way of teaching research skills (Bradstreet, 1996; Marek et al., 2004; Roberts and Allen, 2012, 2013; Earley, 2014; Stoloff et al., 2015). This point was echoed by the academic participants in the current research, who reflected on how embedding statistical decision making in a context of substantive interest, and providing opportunities to work with personally meaningful data promotes student engagement. As an example, in the first author's second year experimental methods and statistics unit, students participate in an experiment early in the semester, which forms the basis of a research report assessment. The topic varies from year to year, but typically involves studying a well established phenomenon in a contemporary context (e.g., the attractiveness stereotype on Facebook; or the Internet as a transactive memory source). In a series of class and homework exercises, students are required to develop one or two theoretically meaningful hypotheses, use the class generated data to test them, and then prepare an American Psychological Association (APA) style research report for assessment. The experiment is usually structured such that several meaningful hypotheses are possible, and testable using techniques taught in the unit (which include parametric and non-parametric tests for comparing independent and related groups). One of the key tasks in this process is the identification of an appropriate statistical test for each hypothesis. Of course, such class research projects need not be the exclusive domain of research methods and statistics units, and can also be deployed effectively to teach a wide range of subjects (e.g., Lutsky, 1986; Ragozzine, 2002).

The second aim of this research was to solicit psychology students' and academics' views on the nature of resources that could facilitate the statistical decision making process. The findings indicate that both groups support the development of a digital decision tree that is simple to use, easy to access, provides multiple levels of depth, and is endorsed by "experts." The psychology academics also stressed the need for such a resource to function as a teaching tool, which engages students with each choice-point in the decision making process, rather than simply providing an "answer." This is in contrast to some recent trends in statistics software development to automate the test selection process based on the characteristics of the user's data file (e.g., "Nonparametric Tests" in IBM SPSS; Wacharamanotham et al., 2015). In fact, such trends are antithetical to the views of the academics in our sample, who strongly believed that statistics should be considered concurrently with other design issues, and far before any data are collected.

Based partially on the findings of the current study, as well as existing literature on the efficacy of decision trees and mobile learning technologies, we have recently published StatHand (see https://stathand.net), a free cross-platform mobile application designed to support students through the statistical decision making process. This application, developed with the support of the Australian Government Office for Learning and Teaching, guides users through a series of annotated questions to ultimately offer them the guidance necessary to conduct a suitable statistical test, as well as interpret and report its results. A full discussion of StatHand is beyond the scope of this paper, but interested readers are referred to Allen et al. (under review). In this paper, we overview the rationale behind StatHand, describe the development process and feature set of the application, and provide guidelines for integrating its use into the research methods curriculum.

When interpreting the findings of this research, readers should give consideration to the usual caveats regarding small samples and the transferability of qualitative research findings. The nature of the task we asked of participants (i.e., to describe how they would identify a suitable statistic) also warrants some consideration. It is plausible that the apparent deftness with which the academics approached this task is at least partially a function of the nature of their work, in which we imagine they routinely practice the metacognition and self-reflection for which we probed[4]. By contrast, it is suspected that the students in the sample have less experience with such skills, and fewer daily opportunities to practice them. However, this is a matter requiring attention in future research. Future research should also focus on exploring theoretically driven strategies and resources that may facilitate the statistical decision making process, and speed up the development of selection skills and structural awareness. To date, work in this area has largely focused on involving students in concrete research projects (e.g., Kardash, 2000) or the use of decision trees (e.g., Carlson et al., 2005; and the current research). Future work should be methodologically rigorous, and based on experimental methods, rather than the non-experimental and quasi-experimental approaches so commonly utilized in teaching and learning research (Wilson-Doenges and Gurung, 2013).

In conclusion, this paper presents a qualitative exploration of the strategies psychology students and academics use to make statistical decisions. The students in our sample found this task challenging, and many struggled to describe a coherent strategy for choosing appropriate statistical tests for common research scenarios. Those who did recognize that such scenarios could be approached in a systematic fashion tended to reflect on the utility of decision trees they had encountered in their studies. Unlike the students, the academics described selecting appropriate statistics as a

---

[4]As kindly noted by one of our reviewers, the apparent deftness with which the academic participants were able to explore possibilities and identify suitable statistics sits in contrast with our discipline's well known difficulties when it comes to interpreting such statistics (e.g., Cohen, 1994; Hoekstra et al., 2006, 2014; McGrath, 2011; Kline, 2013).

complex, nuanced, and iterative process, embedded within the broader process of conducting research. When both groups were asked to imagine tools or resources that could facilitate the statistical decision making process, they tended to describe digital technologies based on a decision-tree framework. To the academics in particular, it was important that such resources scaffold the development of independent decision making competence, and not strip the user of the learning opportunities inherent in working through the full research process.

AUTHOR CONTRIBUTIONS

PA conceived and designed the study with the support of LR. KD conducted the second phase interviews. PA analyzed the data and lead the writing of this manuscript, both with support and contributions from LR and KD.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00188

REFERENCES

Allen, P., Bennett, K., and Heritage, B. (2014). *SPSS Statistics Version 22: A Practical Guide.* Melbourne, VIC: Cengage Learning.

American Psychological Association Board of Educational Affairs Task Force on Psychology Major Competencies (2013). *APA Guidelines for the Undergraduate Psychology Major: Version 2.0.* Washington, DC: American Psychological Association Board of Educational Affairs Task Force on Psychology Major Competencies. Retrieved from http://www.apa.org/ed/precollege/about/psymajor-guidelines.pdf

American Psychological Association Presidential Task Force on Evidence Based Practice (2006). Evidence-based practice in psychology. *Am. Psychol.* 61, 271–285. doi: 10.1037/0003-066X.61.4.271

Australian Psychology Accreditation Council (2014). *Accreditation Standard for Programs of Study in Psychology Eligible for Approval Under the National Registration and Accreditation Scheme for the Health Professions: Second Consultation Draft.* Melbourne,VIC: Australian Psychology Accreditation Council. Retrieved from https://www.psychologycouncil.org.au/Assets/Files/APAC_Accreditation_Standard_for_Programs_of_Study_in_Psychology_2nd_Consultation_Draft_27_May_2014.pdf

Beitz, J. M. (1998). Helping students learn and apply statistical analysis: a metacognitive approach. *Nurse Educ.* 23, 49–51. doi: 10.1097/00006223-199801000-00016

Benjamin, L. T. Jr. (2014). *A Brief History of Modern Psychology, 2nd Edn.* Hoboken, NJ: Wiley Blackwell.

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *Am. Stat.* 50, 69–78.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

British Psychological Society (2015). *Standards for the Accreditation of Undergraduate, Conversion and Integrated Masters Programmes in Psychology (October 2015).* Leicester: British Psychological Society. Retrieved from http://www.bps.org.uk/system/files/Public%20files/PaCT/undergraduate_accreditation_2015_web.pdf

Carlson, M., Protsman, L., and Tomaka, J. (2005). Graphic organizers can facilitate selection of statistical tests. Part 1 - analysis of group differences. *J. Phys. Therapy Educ.* 19, 57–66.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *J. Stat. Educ.* 10. Available online at: http://www.amstat.org/publications/jse/v10n3/chance.html

Cohen, J. (1994). The earth is round ($p <.05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Davis, M., Bolding, G., Hart, G., Sherr, L., and Elford, J. (2004). Reflecting on the experiences of interviewing online: perspectives from the Internet and HIV study in London. *AIDS Care* 16, 944–952. doi: 10.1080/09540120412331292499

Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teach. High. Educ.* 19, 242–253. doi: 10.1080/13562517.2013.860105

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics, 4th Edn.* London: Sage.

Gardner, P. L., and Hudson, I. (1999). University students' ability to apply statistical procedures. *J. Stat. Educ.* 7. Available online at: http://www.amstat.org/publications/jse/secure/v7n1/gardner.cfm

Garfield, J., and Ben-Zvi, D. (2007). How students learn statistics revisited: a current review of research on teaching and learning statistics. *Int. Stat. Rev.* 75, 372–396. doi: 10.1111/j.1751-5823.2007.00029.x

Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006). Probability as certainty: dichotomous thinking and the misuse of *p* values. *Psychon. Bull. Rev.* 13, 1033–1037. doi: 10.3758/BF03213921

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychon. Bull. Rev.* 21, 1157–1164. doi: 10.3758/s13423-013-0572-3

Kardash, C. M. (2000). Evaluation of undergraduate research experience: perceptions of undergraduate interns and their faculty mentors. *J. Educ. Psychol.* 92, 191–201. doi: 10.1037/0022-0663.92.1.191

Kidd, S. A. (2002). The role of qualitative research in psychology journals. *Psychol. Methods* 7, 126–138. doi: 10.1037/1082-989X.7.1.126

Kline, R. B. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences, 2nd Edn.* Washington, DC: American Psychological Association. doi: 10.1037/14136-000

Koch, C., and Gobell, J. (1999). A hypertext-based tutorial with links to the Web for teaching statistics and research methods. *Behav. Res. Methods Instrum. Comput.* 31, 7–13. doi: 10.3758/BF03207686

Lovett, M. C., and Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *Am. Stat.* 54, 196–206. doi: 10.1080/00031305.2000.10474545

Lutsky, N. (1986). Undergraduate research experience through the analysis of data sets in psychology courses. *Teach. Psychol.* 13, 119–122. doi: 10.1207/s15328023top1303_4

Marek, P., Christopher, A. N., and Walker, B. J. (2004). Learning by doing: Research methods with a theme. *Teach. Psychol.* 31, 128–131. doi: 10.1080/03075079512331381703

McGovern, T. V., Corey, L., Cranney, J., Dixon, W. E. Jr., Holmes, J. D., Kuebli, J. E., et al. (2010). "Psychologically literate citizens," in *Undergraduate Education in Psychology: A Blueprint for the Future of the Discipline,* ed D. F. Halpern (Washington, DC: American Psychological Association), 9–27.

McGrath, R. E. (2011). *Quantitative Models in Psychology*. Washington, DC: American Psychological Association. doi: 10.1037/12316-000

Mock, T. J. (1972). A decision tree approach to the methodological decision process. *Account. Rev.* 47, 826–829.

Murtonen, M., Olkinuora, E., Tynjala, P., and Lehtinen, E. (2008). "Do I need research skills in working life?": University students' motivation and difficulties in quantitative methods courses. *High. Educ.* 56, 599–612. doi: 10.1007/s10734-008-9113-9

National Health and Medical Research Council, Australian Research Council and Australian Vice-Chancellors' Committee (2007). *National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015)*. Canberra, ACT: National Health and Medical Research Council, Australian Research Council and Australian Vice-Chancellors' Committee.

Nesbit, J., and Adesope, O. (2006). Learning with concept and knowledge maps: a meta-analysis. *Rev. Educ. Res.* 76, 413–448. doi: 10.3102/00346543076003413

Perlman, B., and McCann, L. I. (1999). The most frequently listed courses in the undergraduate psychology curriculum. *Teach. Psychol.* 26, 177–182. doi: 10.1207/S15328023TOP260303

Perlman, B., and McCann, L. I. (2005). Undergraduate research experiences in psychology: a national study of courses and curricula. *Teach. Psychol.* 32, 5–14. doi: 10.1207/s15328023top3201_2

Protsman, L., and Carlson, M. (2008). Graphic organizers can facilitate selection of statistical tests: Part 2 – Correlation and regression analysis. *J. Phys. Therapy Educ.* 22, 36–41.

Quilici, J. L., and Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Educ. Psychol.* 88, 144–161. doi: 10.1037/0022-0663.88.1.144

Quilici, J. L., and Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Appl. Cogn. Psychol.* 16, 325–342. doi: 10.1002/acp.796

Rabinowitz, M., and Hogan, T. M. (2008). Experience and problem representation in statistics. *Am. J. Psychol.* 121, 395–407. doi: 10.2307/20445474

Ragozzine, F. (2002). SuperLab LT: evaluation and uses in teaching experimental psychology. *Teach. Psychol.* 29, 251–254. doi: 10.1207/S15328023TOP2903_13

Roberts, L. D., and Allen, P. J. (2012). "Student perspectives on the value of research participation," in *Teaching Psychology Around the World: Volume 3*, eds S. McCarthy, K. L. Dickson, J. Cranney, A. Trapp, and V. Karandashev (Newcastle upon Tyne: Cambridge Scholars Publishing), 198–212.

Roberts, L. D., and Allen, P. J. (2013). A brief measure of student perceptions of the educational value of research participation. *Aust. J. Psychol.* 65, 22–29. doi: 10.1111/ajpy.12007

Schau, C., and Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *Am. Stat.* 51, 171–175.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychol. Methods* 18, 572–582. doi: 10.1037/a0034177

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* New York, NY: McGraw-Hill.

Stoloff, M. L., Good, M. R., Smith, K. L., and Brewster, J. (2015). Characteristics of programs that maximize psychology major success. *Teach. Psychol.* 42, 99–108. doi: 10.1177/0098628315569877

Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics, 6th Edn.* Boston, MA: Pearson.

Twycross, A., and Shields, L. (2004). Statistics made simple: part 4–Choosing the right statistical test. *Paediatr. Nurs.* 16:24. doi: 10.7748/paed2004.10.16.8.24.c944

Wacharamanotham, C., Subramanian, K., Völkel, S. T., and Borchers, J. (2015). "Statsplorer: guiding novices in statistical analysis," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 2693–2702. doi: 10.1145/2702123.2702347

Ware, M. E., and Chastain, J. D. (1989). Computer-assisted statistical analysis: a teaching innovation? *Teach. Psychol.* 16, 222–227. doi: 10.1207/s15328023top1604_16

Ware, M. E., and Chastain, J. D. (1991). Developing selection skills in introductory statistics. *Teach. Psychol.* 18, 219–222. doi: 10.1207/s15328023top1804_4

Wilson-Doenges, G., and Gurung, R. A. R. (2013). Benchmarks for scholarly investigations of teaching and learning. *Aust. J. Psychol.* 65, 63–70. doi: 10.1111/ajpy.12011

Yan, J., and Lavigne, N. C. (2014). Promoting college students' problem understanding using schema-emphasizing worked examples. *J. Exp. Educ.* 82, 74–102. doi: 10.1080/00220973.2012.745466

Yin, Y. (2012). Using tree diagrams as an assessment tool in statistics education. *Educ. Assessment* 17, 22–49. doi: 10.1080/10627197.2012.697850

frontiers in Psychology

Introducing StatHand: A Cross-Platform Mobile Application to Support Students' Statistical Decision Making

Peter J. Allen[1]*, Lynne D. Roberts[1], Frank D. Baughman[1], Natalie J. Loxton[2], Dirk Van Rooy[3], Adam J. Rock[4] and James Finlay[1]

[1] School of Psychology and Speech Pathology, Curtin University, Perth, WA, Australia, [2] School of Applied Psychology, Griffith University, Brisbane, QLD, Australia, [3] Research School of Psychology, Australian National University, Canberra, ACT, Australia, [4] School of Behavioural, Cognitive and Social Sciences, University of New England, Armidale, NSW, Australia

Although essential to professional competence in psychology, quantitative research methods are a known area of weakness for many undergraduate psychology students. Students find selecting appropriate statistical tests and procedures for different types of research questions, hypotheses and data types particularly challenging, and these skills are not often practiced in class. Decision trees (a type of graphic organizer) are known to facilitate this decision making process, but extant trees have a number of limitations. Furthermore, emerging research suggests that mobile technologies offer many possibilities for facilitating learning. It is within this context that we have developed StatHand, a free cross-platform application designed to support students' statistical decision making. Developed with the support of the Australian Government Office for Learning and Teaching, StatHand guides users through a series of simple, annotated questions to help them identify a statistical test or procedure appropriate to their circumstances. It further offers the guidance necessary to run these tests and procedures, then interpret and report their results. In this Technology Report we will overview the rationale behind StatHand, before describing the feature set of the application. We will then provide guidelines for integrating StatHand into the research methods curriculum, before concluding by outlining our road map for the ongoing development and evaluation of StatHand.

Keywords: statistics, research methods, selection skills, decision tree, teaching and learning, mobile learning, iOS, web application

INTRODUCTION

Quantitative research methods underpin psychological literacy (McGovern et al., 2010; Cranney and Dunn, 2011; Roberts et al., 2015), and are critical to the development of professional competence in psychology. They have featured prominently in undergraduate psychology curricula since the discipline's formation (Perlman and McCann, 1999; Saville, 2008), and are reflected in the course learning outcomes and graduate attributes specified by accrediting psychology organizations worldwide. For example, the Australian Psychology Accreditation Council [APAC] (2014, p. 35) specify six graduate attributes for an undergraduate psychology program. Two of

these, ("understands the principles of scientific method and is able to apply and evaluate basic research methods in psychology" and "demonstrates the capacity to utilize logic, evidence, and psychological science to evaluate claims about, and solve problems regarding, human behavior"), require a solid and flexible understanding of research methods and statistics. The second of five learning goals for an undergraduate psychology course detailed by American Psychological Association Board of Educational Affairs Task Force on Psychology Major Competencies (2013, p. 15) is "scientific inquiry and critical thinking," which requires "the development of scientific reasoning and problem solving, including effective research methods," "applying research design principles to drawing conclusions about psychological phenomena" and "designing and executing research plans." Similar goals or standards are promoted by the British Psychological Society [BPS] (2014) and other accrediting organizations. Collectively, these standards reflect a widely held understanding that an ability to source, read, understand and critically evaluate relevant research literature is a necessary precursor to evidence-based practice in psychology (American Psychological Association Presidential Task Force on Evidence Based Practice, 2006). The vast majority of this literature is based on quantitative research methods (Kidd, 2002; Rennie et al., 2002). It is also widely held that some of the most effective ways of teaching these skills involve engaging students regularly in all aspects of the research process, from the conception of meaningful research questions, through design, analysis, interpretation and reporting (Marek et al., 2004; Wagner et al., 2011; Earley, 2014; Stoloff et al., 2015). Hence, nearly all psychology departments provide multiple opportunities for undergraduate students to conduct original empirical research, either individually or in collaboration with other students or faculty (Kierniesky, 2005; Perlman and McCann, 2005).

Despite their importance, and their prominence throughout psychology curricula, research methods and (particularly) statistics are recognized areas of weakness for many students (Garfield and Ahlgren, 1988; Murtonen and Lehtinen, 2003; Garfield and Ben-Zvi, 2007; Murtonen et al., 2008). Students are known to particularly struggle with the task of selecting appropriate statistical tests and procedures for different types of research questions, hypotheses and data types; an ability which has been referred to as 'selection skill' (Ware and Chastain, 1989). To illustrate this point, Gardner and Hudson (1999) presented 21 brief research scenarios to a sample of 23 students and asked them to recall appropriate statistical procedures for as many scenarios as possible within a 45-min period. The scenarios reflected statistical concepts typically found in introductory statistics textbooks and widely used in behavioral science research. Despite most students having completed at least six research methods and statistics units[1], they overwhelmingly found the task difficult and performed poorly. On average, students managed to read 10.9 scenarios within the allocated time, and answered 25.3% of them correctly. An additional 15.7% of answers were coded as 'partially correct.' When Gardner and Hudson questioned the students

about how they made their decisions, several explanations for the poor performance emerged. These included students misinterpreting the research scenarios, knowing but being unable to name appropriate statistics, misidentifying the measurement levels (e.g., nominal, ordinal, continuous) of variables, and seizing on misleading keywords and data presentation formats.

When Allen et al. (2016) presented similar research scenarios to undergraduate psychology students, they also found the the task of identifying appropriate statistical tests and procedures particularly challenging. Many were apologetic, and expressed embarrassment at being unable to successfully complete a task they felt they ought to be equipped to accomplish. When prompted to think about the process of selecting a statistical procedure (rather than actually identifying one), they continued to struggle. The processes they described tended to be haphazard and inefficient, and included looking for clues in the wording of scenarios, searching through textbooks, relying on memory or simply guessing. Of those who recognized that a systematic decision making process could be followed; none could identify every factor that would require consideration, and most also focused on irrelevant or peripheral aspects of the scenarios.

When students are asked to recognize (rather than recall) appropriate statistics, their performance appears similarly underwhelming. For example, Ware and Chastain (1989, p. 225) developed an eight-item multiple-choice selection skill test, which they and colleagues believed contained "problems that students should be able to solve after completing [an] introductory statistics course." When they administered the test to students at the conclusion of such a course, the students answered fewer than 45% of the items correctly. Ware and Chastain (1989, p. 226) attributed this poor performance, at least partially, to a curriculum which taught statistical techniques "one at a time," and did not emphasize the development of selection skills. A number of other researchers have also recognized that having relatively few opportunities to practice selection skills could account for the difficulties that students experience when placed in situations where they must work out *which statistic* to use (e.g., Quilici and Mayer, 1996, 2002; Lovett and Greenhouse, 2000; Yan and Lavigne, 2014).

Although not many research methods and statistics courses appear to do so, it is possible to train selection skills. For example, when Ware and Chastain (1991) restructured their introductory statistics course to place greater emphasis on when to use various statistics, and less on computational procedures, they observed a significant improvement on their multiple-choice selection skill test. In a more controlled context, Quilici and Mayer (2002) demonstrated that it is possible to train students to focus on the structural (e.g., the nature of the independent and dependent variables, and the relationship between them) rather than surface-level (e.g., topic) features of basic research scenarios, and that doing so improved students' abilities to correctly categorize scenarios according to how they would be analyzed. After training, students were also better able to generate new scenarios that could be analyzed using the same statistical procedures as existing scenarios. More recently, similar findings were reported by Yan and Lavigne (2014), who observed that providing students with worked examples emphasizing the structural features of

---

[1]In Australia, a 'unit' is a single subject, typically taken alongside two or three others over a semester. This term is analogous to 'course' in the United States.

simple research scenarios improved students' performance on subsequent categorization tasks, as well as their ability to identify the structural features defining each category.

Together, these findings suggest that selection skills are underpinned by 'structural awareness' (Quilici and Mayer, 2002), which reflects an ability to disregard the surface features of a research scenario, and focus on its structural features and the relations between them. Like the worked examples used by Yan and Lavigne (2014), graphic organizers, particularly decision trees and flow charts, provide a pedagogical tool for systematically focusing attention on these structural features and relations.

GRAPHIC ORGANIZERS

Graphic organizers are known to facilitate the process of selecting appropriate statistical tests and procedures for different types of research questions and data. They focus the user on each structural component of a research scenario, and illustrate their connectedness/differentiation with spatial positioning and lines (Nesbit and Adesope, 2006). The structured nature of graphic organizers can help users organize new information and integrate it with existing knowledge into schemata (Yin, 2012). The grouping of information lessens cognitive load, and thus more working memory can be applied to learning and problem solving (Yin, 2012). Furthermore, graphic organizers encourage both verbal and spatial encoding of new information, thus providing multiple pathways for its later recall (Katayama and Robinson, 2000). Meta-analyses support the efficacy of concept maps, a type of graphic organizer, for increasing student achievement (Horton et al., 1993), knowledge retention and transfer (Nesbit and Adesope, 2006), and learning (Moore and Readence, 1984).

A number of different types of graphic organizers have been created to help students select appropriate statistical analyses, including tip sheets which sort analyses by their defining characteristics (e.g., Twycross and Shields, 2004), and charts which link statistics to common research goals (e.g., Beitz, 1998). However, the organizers which have gained most traction follow decision tree logic, and are designed to guide the user from an initial question (or problem) to an answer or outcome, via a series of choice or decision points. In domains that involve complex rules, procedures, conditions, and multiple candidate solutions, the use of a decision tree can provide a highly organized approach to the process of decision-making. In the domain of statistics, decision-trees to guide statistical decision making have a long history (e.g., Mock, 1972; Fok et al., 1995) and are now commonly included in statistics textbooks (see, for e.g., Tabachnick and Fidell, 2013; Allen et al., 2014). Statistical decision trees differ from other types of graphic organizers in that they are hierarchical and start with a single node before branching off. By following the branches that refer to the key structural details of a research scenario, the user is led to a statistical analysis appropriate to their circumstances (Mertler and Vannatta, 2002). Theoretically, decision trees rest on the idea that knowledge must be organized or structured to be accessible from long-term memory (Schau and Mattern, 1997). Decision trees provide this structure by explicitly highlighting the interconnectedness (and

differentiation) between important statistical concepts (Schau and Mattern, 1997; Yin, 2012).

Empirically, there is work illustrating both the objective efficacy of statistical decision trees, as well as their subjective appeal. For example, Carlson et al. (2005; Protsman and Carlson, 2008) demonstrated that decision trees could facilitate significantly faster and more accurate (by a multiple of three) statistical decision-making, compared to more traditional methods of statistical test selection (e.g., by searching through a familiar textbook). The decision tree method was also significantly more popular amongst students than the textbook method (Carlson et al., 2005; Protsman and Carlson, 2008).

Despite their popularity, traditional statistical decision trees also have limitations. First, they are usually limited in scope by the requirement to fit them on a single sheet of paper, or within the pages of a textbook. Consequently, definitions and other information that would make traversing the tree easier are either spatially separated from the tree itself, or completely absent (Koch and Gobell, 1999; Blankenship and Dansereau, 2000). Second, when given to students without accompanying resources (e.g., a textbook) they do not provide sufficient detail to execute and interpret the statistics they help identify. Third, while the complexity and non-linearity of a statistical decision tree may be helpful to experienced users, new users may experience difficulty in fully processing the tree (sometimes referred to as 'map shock'), and consequently lose the motivation to use it (Blankenship and Dansereau, 2000; Nesbit and Adesope, 2011).

To overcome these limitations, a number of researchers and educators have adapted the traditional decision tree model for digital media. These hypertext systems are typically comprised of a series of interconnected pages or nodes (Unz and Hesse, 1999). Space constraints associated with paper decision trees are removed, and links can be made to external resources that aid learning (Koch and Gobell, 1999). Map shock can be eliminated because the user is only shown a small section of the tree at any given time, reducing its complexity and ability to overwhelm (Blankenship and Dansereau, 2000). However, a hypertext system can provide a disjointed experience, where users become disoriented and lose track of their location within the system. This phenomenon, sometimes referred to as 'lost in hyperspace' (Otter and Johnson, 2000), can constrain the novice user's ability to develop an understanding of how concepts are connected. Despite this limitation, meta-analytic findings support the overall efficacy of hypertext systems in comparison to textual interfaces. In particular, when compared to textual interfaces, graphical map interfaces are associated with more effective (medium to large effect sizes) and efficient (small to medium effect sizes) performance (Chen and Rada, 1996).

Koch and Gobell (1999) adapted paper decision trees for delivery on the world-wide-web, and in doing so were able to also provide users with definitions, links to online resources, and information about how to enter and analyze data in commonly used statistical software. Like Carlson et al. (2005; Protsman and Carlson, 2008), Koch and Gobell (1999) found that students using their online decision tree were better able to identify appropriate statistical tests than students in a comparison condition. Unfortunately, Koch and Gobell's (1999) website is

no longer active. A current example of an online statistical test selection tool is that provided by University of California, Los Angeles (UCLA)'s Institute for Digital Research and Education at http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm. This site provides a table of statistical tests based on the number and nature of dependent and independent variables, with 'how to' links for a range of statistical software. However, the large size of the table (and the use of a table rather than a decision tree format) combined with the limited information provided may contribute to map-shock for inexperienced users.

A range of software for selecting statistical techniques has also been developed. Some software applications currently available (e.g., Subramanian, 2014; Wacharamanotham et al., 2015) automatically select the statistical test for the user without explicitly guiding the user through the steps to make the decision, greatly reducing their pedagogic potential. STestMAP (Eng et al., 2011) is a visual tool that guides students through a systematic process to select a statistical test, but does not appear to be publicly available. Despite their potential benefits, hypertext decision trees and currently available software generally require the user to have a live internet connection.

MOBILE LEARNING TECHNOLOGIES

Unlike websites and web applications, mobile learning applications can be developed to maintain all (or most) of their functionality in the absence of an internet connection (Kretser et al., 2015). Mobile learning can be defined as "the use of mobile or wireless devices for the purpose of learning while on the move" (Park, 2011, cited in Yu et al., 2014, p. 2126). In the previous decade, the use of mobile learning technologies such as smart devices and mobile applications has increased rapidly, and amongst western higher education students their penetration is near ubiquitous (Stowell, 2011; Murphy et al., 2013; Dahlstrom and Bichsel, 2014; Chen et al., 2015). Their broad appeal is tied to many factors, including portability, enabling the user to access information and resources virtually anywhere and at any time (Jeng et al., 2010), and utility. Increasingly, students prefer to use their own smart devices for learning, and mobile learning applications have been identified as one of the technologies expected to have the biggest impact on education this decade (Martin et al., 2011; Johnson et al., 2012). In the context of teaching research methods and statistics, emerging research suggests that technology assisted examples delivered via mobile applications positively impact on student learning (Harnish et al., 2012).

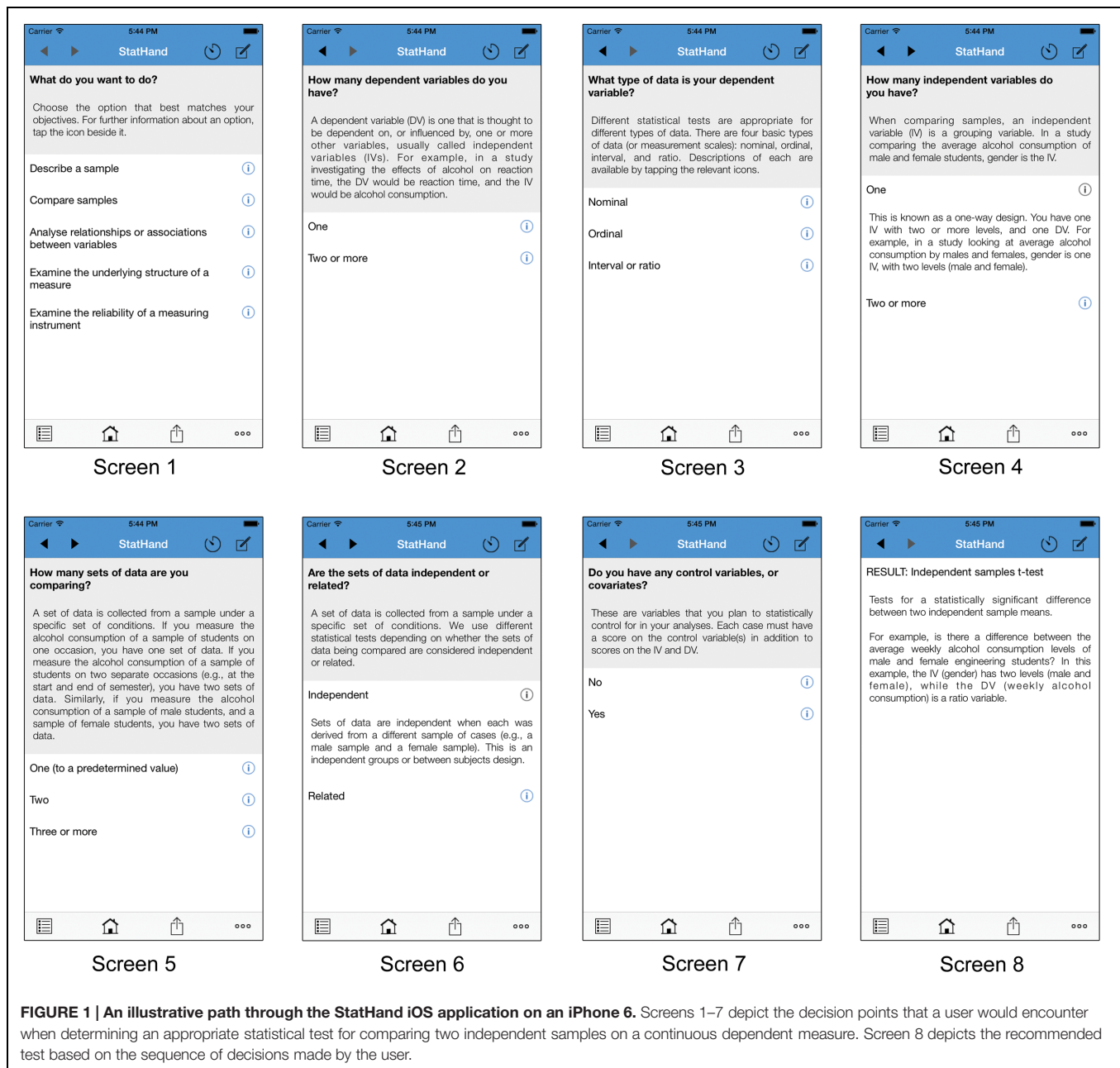## STATHAND: A MOBILE APPLICATION TO SUPPORT STATISTICAL DECISION MAKING

In the previous sections of this paper, we have described how students find statistical test selection difficult, argued that decision trees can facilitate this decision making process, and

noted the rapid adoption of smart devices and mobile learning applications in the higher education sector. With these points in mind, we proposed StatHand to the Australian Government Office for Learning and Teaching in 2013. StatHand was described as a cross-platform mobile application that helps users quickly identify appropriate statistical tests and analytic procedures for a wide range of research questions, hypotheses and data types. The proposal, to develop, disseminate and evaluate StatHand, was funded.

The content of StatHand is being developed in two main phases. The first phase, which is now complete, is focused on helping users identify statistical tests and procedures appropriate to a wide range of circumstances. It is freely available in the iOS App Store, and can also be accessed as a fully mobile-compatible web application at https://stathand.net. The second phase, which is currently under development, guides the computation, interpretation and reporting of these tests and procedures.

The first phase of content is illustrated in **Figure 1**, on the iOS iPhone application. When StatHand is launched (Screen 1), the user is presented with the first of several annotated questions, "what do you want to do?" There are five options available: 'describe a sample,' 'compare samples,' 'analyze relationships or associations between variables,' 'examine the underlying structure of a measure,' and 'examine the reliability of a measuring instrument.' The statistics, tests and procedures under each of these objectives are listed in **Table 1**. Let's imagine that we are planning a simple study to examine whether caffeine affects response time. Response time data will be collected for two groups of adults, who will drink either coffee or water immediately prior to testing. The most appropriate option on Screen 1 is 'compare samples,' as we wish to compare the performance of the coffee drinkers with that of the water drinkers. After making our first selection, we are presented with a second choice, in which we need to identify the number of dependent variables in the study. A user uncertain about what is meant by 'dependent variable' can consult the brief annotation below the question, whereas more experienced users can simply make their selection. Here, we indicate that we have 'one' dependent variable (Screen 2), which is measured on an 'interval or ratio' scale (Screen 3). Next, we are promoted to consider the number and nature of our independent variable(s). As illustrated in Screens 4 and 6, each option can be expanded for context-specific definitions and examples by tapping on the relevant Information icons. Finally, we are asked to indicate whether or not we have any control variables (Screen 7) which, in the current example, we do not. Having now engaged with each relevant structural feature of our research scenario, we are presented with an appropriate analytic choice (Screen 8). In this case, an independent samples *t*-test.

At any point during the decision making process, a user can review their previous choices using the History tool, as illustrated in Screen 9 of **Figure 2**. This feature allows the user to retrace their steps, and draw stronger connections between their choices and the solutions they reach. Selecting any entry

FIGURE 1 | An illustrative path through the StatHand iOS application on an iPhone 6. Screens 1–7 depict the decision points that a user would encounter when determining an appropriate statistical test for comparing two independent samples on a continuous dependent measure. Screen 8 depicts the recommended test based on the sequence of decisions made by the user.

in the History returns the user to the corresponding decision point. Users can also navigate through StatHand using the Back and Forward buttons, or jump directly to a statistic from the searchable Index (illustrated in Screen 10). Also illustrated in Screen 9, **Figure 2** is the Notes tool, with which the user can pin their own annotations to specific pages within the application, or retrieve notes made on other pages. Finally, tapping on the Share icon in the toolbar at the bottom of the screen reveals options to print, email or save the annotated sequence of decisions leading to the current page (including the Notes associated with those decisions). It should be noted that these features work in comparable ways in the web version of StatHand at https://stathand.net, which has been designed for

compatibility with any device capable of running a modern web browser.

SUGGESTIONS FOR INTEGRATING STATHAND INTO THE RESEARCH METHODS CURRICULUM

As we've observed, many psychology students find the task of selecting appropriate statistics for different research questions, hypotheses and data types challenging (Gardner and Hudson, 1999; Allen et al., 2016). This selection skill (Ware and Chastain, 1989) appears underpinned by structural awareness (Quilici

TABLE 1 | The statistics, tests and procedures described in StatHand, grouped by research objective.

Objective	Statistics, tests and procedures described in StatHand
Describe a sample	Bar graph; category count; histogram; interquartile range; Mean; median; mode; pie chart; range; standard deviation; stem-and-leaf plot.
Compare samples	ANCOVA (independent samples and mixed; one way and factorial); ANOVA (independent samples, repeated measures and mixed; one way and factorial); chi-square (goodness of fit and contingencies); Cochran's Q test; Friedman two-way ANOVA; Kruskal–Wallis one-way ANOVA; Mann–Whitney U test; McNemar test of change; t-test (one sample, independent samples and paired samples); Wilcoxon signed-rank test (one sample and paired samples).
Analyze relationships or associations between variables	Chi-square test of contingencies (with Phi or Cramer's V); correlation coefficients (point-biserial, rank-biserial, Spearman's and Pearson's); eta; linear regression (bivariate and multiple; standard and hierarchical); logistic regression (binary and multinomial; standard and hierarchical); ordinal regression (standard and hierarchical).
Examine the underlying structure of a measure	Confirmatory factor analysis; exploratory factor analysis; principal components analysis.
Examine the reliability of a measuring instrument	Cohen's kappa; Cronbach's alpha; intraclass correlation; Kuder–Richardson 20; Weighted kappa.

The objectives listed correspond with the five options presented to users on the StatHand home screen.

and Mayer, 2002, p. 326); an ability to disregard the surface features of research scenarios, and instead focus on their structural features and the relations between them. Traditional research methods and statistics courses underemphasize these skills, although research suggests that they can be trained (e.g., Quilici and Mayer, 2002; Yan and Lavigne, 2014). Decision trees provide a pedagogic tool for systematically focusing attention on the structural features of research scenarios, as well as the relations between them. StatHand reflects a new breed of interactive decision tree, ready for embedding in existing research methods and statistics curricula. It can be used to provide novel and engaging opportunities to practice selection skills and train students' structural awareness by systematically sensitizing them to the issues that require consideration before choosing between statistical techniques. Once the second phase content has been deployed, it can further be used as an aid to guide their computation, interpretation and reporting.

Research suggests that integrating technology generally (e.g., Tishkovskaya and Lancaster, 2012; Moreau, 2015), and mobile applications specifically (Harnish et al., 2012) into the research methods and statistics classroom can have pedagogic benefits. However, doing so is not without challenges. Potential barriers to successful integration include the limited confidence of teachers and students when working with new technologies, and differences in learning and teaching styles. Importantly, Lahiri and Moseley (2012, p. 11) cautioned that the use of smart devices as eLearning tools must be underpinned by pedagogical principles and an evidence base, otherwise the use of such

tools "might lead to frustration, inequity, shallow learning, and distraction from the main purpose of enhancing learning and making students competent professionals." Thus, in order to reduce students' statistics anxiety and enhance students' selection skills, teachers may wish to consider carefully how to effectively use smart devices as part of the learning process. Yu et al. (2014) stress that smart devices need to be used to extend the reach of teaching. Consequently, "shifting from e-learning to mobile learning implies that instructional designers need to adopt new ways of facilitating learning, not in one way, but using multiple pedagogical strategies, to help people learn whenever they need and wherever they are" (Yu et al., 2014, p. 2132).
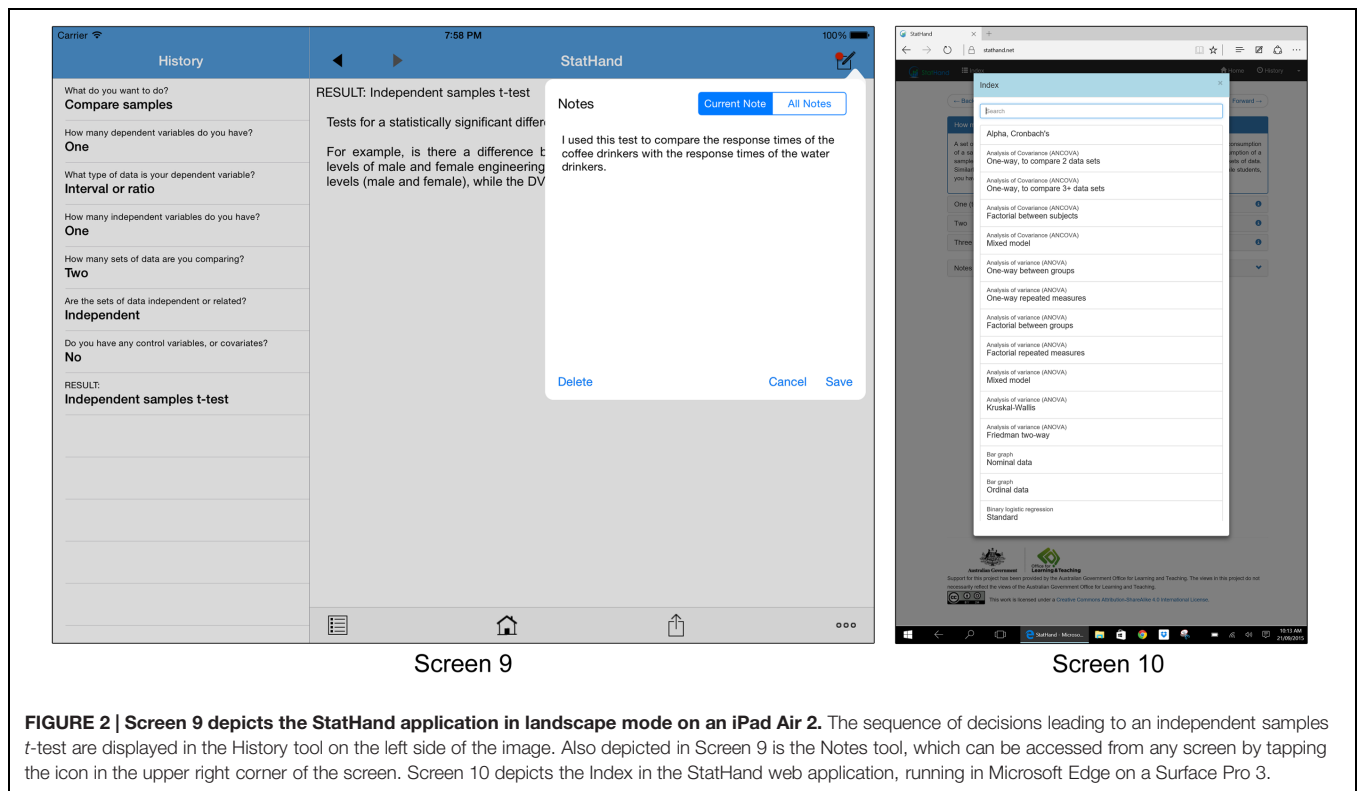
StatHand was developed within the theoretical framework of the Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003). This theory posits that performance expectancy, effort expectancy, social influence and facilitating conditions are direct determinants of the intention to use a particular technology, with intention and facilitating conditions predictors of actual use. Below we offer some suggestions for embedding StatHand in research methods and statistics courses.

Demonstrate StatHand at the Outset and Throughout the Course

StatHand is easily and freely accessible via the iOS App Store and online at https://stathand.net. Navigation through the application is intuitive (although brief instructions are available within the application), and largely self-contained, with definitions and examples of all key terms available at a simple tap of an icon. These features increase effort expectancy (defined in terms of ease of use, Venkatesh et al., 2003) Nevertheless, to maximize the application's perceived utility to students (part of performance expectancy), instructors should devote class time early in the semester to demonstrating how and when to use it. Revisiting StatHand each time a new analysis is introduced will help sensitize students to the similarities and differences between tests vis-à-vis their key structural characteristics (e.g., the key structural difference between the independent samples t-test and ANOVA is the number of levels of the independent variable). Such sensitivity is key to structural awareness, and the development of selection skills. Some instructors already use traditional (paper based) decision trees in efforts to achieve this aim. The benefits of transitioning to StatHand include the reduced potential for map-shock or 'glossing over key decision points,' the provision of an additional set of examples that students can refer to when seeking to master complex concepts, and the ability for students to save, print or email a record of their sequence of decisions (and annotations associated with those decisions) for later reference. Performance expectancy will increase as students succeed in selecting appropriate statistical techniques using StatHand.

Link StatHand to Existing Teaching Resources

StatHand can be easily incorporated into existing teaching activities and resources. For example, one of us (NL) created

FIGURE 2 | Screen 9 depicts the StatHand application in landscape mode on an iPad Air 2. The sequence of decisions leading to an independent samples *t*-test are displayed in the History tool on the left side of the image. Also depicted in Screen 9 is the Notes tool, which can be accessed from any screen by tapping the icon in the upper right corner of the screen. Screen 10 depicts the Index in the StatHand web application, running in Microsoft Edge on a Surface Pro 3.

a YouTube screencast demonstrating the use of StatHand and embedded a link to the screencast (along with links to StatHand) in an existing worksheet demonstrating how to perform and interpret a specific statistical procedure. Another of us (PA) regularly uses it in tutorial activities and assessments, where students are presented with a research scenario and data set, and required to generate meaningful hypotheses. StatHand is then used to identify appropriate hypothesis tests, which are conducted and interpreted in the remainder of the class. The linking of StatHand to existing teaching resources combined with the annotated question feature of the StatHand app provide organization and technical infrastructure (facilitating conditions) to support adoption and use. The use of StatHand within existing forums such as discussion boards and social media sites facilitates social influence, particularly if used across multiple courses within the student's degree.

Minimize Competition from other Sources

Competition from other sources of interaction when using technology in the classroom can impact on focus. To limit such distractions, students will need to be given clear advice about how to maximize the benefits that can be derived from using learning technologies. At a minimum, this may include recommending turning on 'airplane' mode on smart devices, which will prevent them from receiving notifications, and reduce students' temptation to check emails, browse the web or use social networking applications.

Use StatHand Consistently and Repeatedly Throughout the Course (and other Related Courses)

When used effectively, StatHand can reinforce information provided by instructors, and offer practical experience in determining appropriate analyses for a variety of different research scenarios. When used consistently through statistics courses, and when statistical decision-making is explicitly assessed, selection skills can be generalized to other research-related courses. As a single application available free on a wide variety of platforms, StatHand can be readily incorporated across multiple courses in statistics and other research-focused courses throughout the psychology undergraduate degree. Over time, students will become increasingly familiar with StatHand, the promotion of its use by multiple instructors will enhance social influence, and both the intention to use, and actual use of StatHand. Its use will be second nature by the time they begin conducting individual (or small group) research projects in their final years of study.

FUTURE DIRECTIONS AND CONCLUSION

StatHand is a cross-platform application designed to aid the process of selecting statistical tests and procedures for a wide range of research scenarios. It is currently available in the iOS App Store and at https://stathand.net. StatHand can be easily integrated into existing teaching and learning activities, or used

as a base for the development of new activities focused on exploring the circumstances in which different statistics are appropriate.

Content for the second phase of StatHand is currently under development. When incorporated into the iOS and web applications, it will guide users through the computation, interpretation and reporting of each statistic that StatHand helps identify (see **Table 1**). It will also provide advice on testing assumptions and calculating and interpreting effect sizes where appropriate; offer links to additional reputable information about each technique; and highlight controversies and alternative approaches where applicable. Much of this material is being prepared as short videos, developed following evidence-based multimedia learning object design principles (e.g., Clark and Mayer, 2011).

We have also started integrating StatHand into our own research methods and statistics units. This is informing the development of a set of instructors' resources to complement StatHand. These resources will include a brief rationale for the use of the application as a learning and teaching tool, instructions for using the application, tips for integrating StatHand into undergraduate research methods and statistics classes, and active learning activities that instructors can adapt for their own teaching purposes. The package of activities will include multiple-choice quizzes that instructors can use to assess their students' abilities to identify appropriate statistical tests and procedures under a wide variety of circumstances. These will be provided in formats suitable for inclusion in worksheets and tests, as well as formats suitable for inclusion in PowerPoint presentations that either do or do not make use of common audience response technologies (e.g., Turning Point Keepad). When available, the StatHand instructors' resources will be provided freely, on request, to anyone who teaches research methods, statistics and related subjects at recognized higher education institutions.

Dissemination of StatHand is ongoing, and as its user base expands we are collecting usage data that will inform how the application may be optimized to facilitate learning and the decision making process. Additional research projects are experimentally investigating the instructional efficiency of StatHand relative to other common decision making aids (e.g.,

paper based decision trees and familiar textbooks). Further research will empirically investigate students' adoption and use of StatHand within the Unified Theory of Acceptance and Use of Technology framework (Venkatesh et al., 2003). Finally, we will soon begin investigating how instructors use StatHand to support the learning and teaching within their own courses. This multi-pronged evaluation approach has two ultimate aims. The first of these is to inform the ongoing development of StatHand. The second is to develop an evidence base and best-practice recommendations to guide its use.

To conclude, in this Technology Report we have provided an overview of StatHand, a free cross-platform mobile application designed to support students' statistical decision making. Developed with the support of the Australian Government Office for Learning and Teaching, StatHand guides users through a series of simple, annotated questions to help them identify a statistical test or procedure appropriate to their circumstances. In its next release, StatHand will also guide the computation, interpretation and reporting of the tests and procedures it helps users identify. We invite psychology research methods and statistics instructors to contact us about incorporating StatHand into their own classes.

AUTHOR CONTRIBUTIONS

PA led the development of the StatHand application, with support and contributions from LR, FB, NL, DVR, and AR. All authors contributed to the preparation of this manuscript.

ACKNOWLEDGMENTS

REFERENCES

Allen, P., Bennett, K., and Heritage, B. (2014). *SPSS Statistics Version 22: A Practical Guide*. Melbourne, Australia: Cengage Learning.

Allen, P. J., Dorozenko, K. P., and Roberts, L. D. (2016). Difficult decisions: a qualitative exploration of the statistical decision making process from the perspectives of psychology students and academics. *Front. Psychol.* 7:188. doi: 10.3389/fpsyg.2016.00188

American Psychological Association Board of Educational Affairs Task Force on Psychology Major Competencies (2013). *APA Guidelines for the Undergraduate Psychology Major: Version 2.0*. Washington, DC: American Psychological Association Board of Educational Affairs Task Force on Psychology Major Competencies.

American Psychological Association Presidential Task Force on Evidence Based Practice (2006). Evidence-based practice in

psychology. *Am. Psychologist* 61, 271–285. doi: 10.1037/0003-066X.61.4.271

Australian Psychology Accreditation Council [APAC] (2014). *Accreditation Standard for Programs of Study in Psychology Eligible for Approval Under the National Registration and Accreditation Scheme for the Health Professions: Second Consultation Draft*. Melbourne: APAC.

Beitz, J. (1998). Helping students learn and apply statistical analysis: a metacognitive approach. *Nurse Educ.* 23, 49–51. doi: 10.1097/00006223-199801000-00016

Blankenship, J., and Dansereau, D. (2000). The effect of animated node-link displays on information recall. *J. Exp. Educ.* 68, 293–308. doi: 10.1080/00220970009600640

British Psychological Society [BPS] (2014). *Standards for the Accreditation of Undergraduate, Conversion and Integrated Masters Programmes in Psychology*. Leicester: BPS.

Carlson, M., Protsman, L., and Tomaka, J. (2005). Graphic organizers can facilitate selection of statistical tests. *Part 1 - analysis of group differences. J. Phys. Ther. Educ.* 19, 57–66.

Chen, B., Sellhamer, R., Bennett, L., and Bauer, S. (2015). *Students' Mobile Learning Practices in Higher Education: A Multi-Year Study. Educause Review.* Available at: http://er.educause.edu/articles/2015/6/students-mobile-learning-practices-in-higher-education-a-multiyear-study

Chen, C., and Rada, R. (1996). Interacting with hypertext: a meta-analysis of experimental studies. *Hum. Comput. Interact.* 11, 125–156. doi: 10.1207/s15327051hci1102_2

Clark, R. C., and Mayer, R. E. (2011). *E-Learning and the Science of Instruction*, 3rd Edn. San Francisco, CA: Pfeiffer.

Cranney, J., and Dunn, D. S. (2011). Psychological literacy and the psychologically literate citizen: new frontiers for a global discipline," *The Psychologically Literate Citizen: Foundations and Global Perspectives,* eds J. Cranney and D. S. Dunn (Oxford: Oxford University Press), 3–12.

Dahlstrom, E., and Bichsel, J. (2014). *ECAR Study of Undergraduate Students and Information Technology, 2014*. Louisville, CO: Educause Center for Analysis and Research.

Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teach. High. Educ.* 19, 242–253. doi: 10.1080/13562517.2013.860105

Eng, T. H., Hua, L. U., Kheng, F. K., Ying, L. C., and Yee Hui, R. C. (2011). Significance test selection via STestMAP. *GSTF Business Rev.* 1, 155–160. doi: 10.5176_2010-4804_1.2.69

Fok, L. Y., Angelidis, J. P., Ibrahim, N. A., and Fok, W. M. (1995). A decision tree approach to the interpretation of multivariate statistical techniques. *J. Educ. Business* 71, 110–117. doi: 10.1080/08832323.1995.10116770

Gardner, P. L., and Hudson I. (1999). University students' ability to apply statistical procedures. *J. Stat. Educ.* 7.

Garfield, J., and Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *J. Res. Math. Educ.* 19, 44–63. doi: 10.2307/749110

Garfield, J., and Ben-Zvi, D. (2007). How students learn statistics revisited: a current review of research on teaching and learning statistics. *Int. Stat. Rev.* 75, 372–396. doi: 10.1111/j.l751-5823.2007

Harnish, D., Ling, C., and Shehab, R. (2012). Leveraging the use of mobile applications to increase knowledge retention in a classroom lecture. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.* 56, 610–614. doi: 10.1177/1071181312561127

Horton, P., Mcconney, A., Gallo, M., Woods, A., Senn, G., and Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. *Sci. Educ.* 77, 95–111. doi: 10.1002/sce.3730770107

Jeng, Y. L., Wu, T. T., Huang, Y. M., Tan, Q., and Yang, S. J. H. (2010). The add-on impact of mobile applications in learning strategies: a review study. *Educ. Technol. Soc.* 13, 3–11.

Johnson, L., Adams, S., and Cummins, M. (2012). *Technology Outlook for Australian Tertiary Education 2012–2017: An NMC Horizon Report Regional Analysis*. Austin, TX: The New Media Consortium.

Katayama, A., and Robinson, D. (2000). Getting students "partially" involved in note-taking using graphic organizers. *J. Exp. Educ.* 68, 119–133. doi: 10.1080/00220970009598498

Kidd, S. A. (2002). The role of qualitative research in psychology journals. *Psychol. Methods* 7, 126–138. doi: 10.1037/1082-989X.7.1.126

Kierniesky, N. C. (2005). Undergraduate research in small psychology departments: two decades later. *Teach. Psychol.* 32, 84–90. doi: 10.1207/s15328023top3202_1

Koch, C., and Gobell, J. (1999). A hypertext-based tutorial with links to the Web for teaching statistics and research methods. *Behav. Res. Methods Instrum. Comput.* 31, 7–13. doi: 10.3758/BF03207686

Kretser, H., Wong, R., Roberton, S., Pershyn, C., Huang, J., Sun, F., et al. (2015). Mobile decision-tree tool technology as a means to detect wildlife crimes and build enforcement networks. *Biol. Conserv.* 189, 33–38. doi: 10.1016/j.biocon.2014.08.018

Lahiri, M., and Moseley, J. L. (2012). Is mobile learning the future of 21st century education? Educational considerations from various perspectives. *Educ. Technol.* 52, 3–13.

Lovett, M. C., and Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *Am. Statis.* 54, 196–206. doi: 10.2307/2685590

Marek, P., Christopher, A. N., and Walker, B. J. (2004). Learning by doing: research methods with a theme. *Teach. Psychol.* 31, 128–131. doi: 10.1207/s15328023top3102_6

Martin, S., Diaz, G., Sancristobal, E., Gil, R., Castro, M., and Peire, J. (2011). New technology trends in education: seven years of forecasts and convergence. *Comput. Educ.* 57, 1893–1906. doi: 10.1016/j.compedu.2011.04.003

McGovern, T. V., Corey, L., Cranney, J., Dixon, W. E. Jr., Holmes, J. D., Kuebli, J. E., et al. (2010). "Psychologically literate citizens," *Undergraduate Education in Psychology: A Blueprint for the Future of the Discipline*, ed. D F. Halpern (Washington, DC: American Psychological Association), 9–27.

Mertler, C., and Vannatta, R. (2002). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*, 2nd Edn. Glendale, CA: Pyrczak.

Mock, T. J. (1972). A decision tree approach to the methodological decision process. *Account. Rev.* 47, 826–829.

Moore, D. W., and Readence, J. F. (1984). A quantitative and qualitative review of graphic organizer research. *J. Educ. Res.* 78, 11–17. doi: 10.1080/00220671.1984.10885564

Moreau, D. (2015). When seeing is learning: dynamic and interactive visualizations to teach statistical concepts. *Front. Psychol.* 6:342. doi: 10.3389/fpsyg.2015.00342

Murphy, A., Farley, H., and Koronios, A. (2013). "Understanding the use of smart mobile technologies for learning in higher education," in *Proceedings of Ascilite 2013 Sydney on Electric Dreams*, eds H. Carter, M. Gosper, and J. Hedberg, (Sydney: Ascilite shaping the future of tertiary education), 18–22.

Murtonen, M., and Lehtinen, E. (2003). Difficulties experienced by education and sociology students in quantitative methods courses. *Stud. High. Educ.* 28, 171–185. doi: 10.1080/0307507032000058064

Murtonen, M., Olkinuora, E., Tynjala, P., and Lehtinen, E. (2008). "Do I need research skills in working life?": University students' motivation and difficulties in quantitative methods courses. *High. Educ.* 56, 599–612. doi: 10.1007/s10734-008-9113-9

Nesbit, J., and Adesope, O. (2006). Learning with concept and knowledge maps: a meta-analysis. *Rev. Educ. Res.* 76, 413–448. doi: 10.3102/00346543076003413

Nesbit, J., and Adesope, O. (2011). Learning from animated concept maps with concurrent audio narration. *J. Exp. Educ.* 79, 209–230. doi: 10.1080/00220970903292918

Otter, M., and Johnson, H. (2000). Lost in hyperspace: metrics and mental models. *Interact. Comput.* 13, 1–40. doi: 10.1016/S0953-5438(00)00030-8

Perlman, B., and McCann, L. I. (1999). The most frequently listed courses in the undergraduate psychology curriculum. *Teach. Psychol.* 26, 177–182. doi: 10.1207/S15328023TOP260302

Perlman, B., and McCann, L. I. (2005). Undergraduate research experiences in psychology: a national study of courses and curricula. *Teach. Psychol.* 32, 5–14. doi: 10.1207/s15328023top3201_2

Protsman, L., and Carlson, M. (2008). Graphic organizers can facilitate selection of statistical tests: Part 2 – Correlation and regression analysis. *J. Phys. Ther. Educ.* 22, 36–41.

Quilici, J. L., and Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Educ. Psychol.* 88, 144–161. doi: 10.1037/0022-0663.88.1.144

Quilici, J. L., and Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Appl. Cogn. Psychol.* 16, 325–342. doi: 10.1002/acp.796

Rennie, D. L., Watson, K. D., and Monteiro, A. M. (2002). The rise of qualitative research in psychology. *Can. Psychol.* 43, 179–189. doi: 10.1037/h0086914

Roberts, L. D., Heritage, B., and Gasson, N. (2015). The measurement of psychological literacy: a first approximation. *Front. Psychol.* 6:105. doi: 10.3389/fpsyg.2015.00105

Saville, B. K. (2008). *A Guide to Teaching Research Methods in Psychology*. Malden, MA: Blackwell.

Schau, C., and Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *Am. Statistic.* 51, 171–175. doi: 10.1080/00031305.1997.10473955

Stoloff, M. L., Good, M. R., Smith, K. L., and Brewster, J. (2015). Characteristics of programs that maximize psychology major success. *Teach. Psychol.* 42, 99–108. doi: 10.1177/0098628315569877

Stowell, J. R., (2011). "Emerging technologies to improve teaching and learning in a digital world," in *Best Practices for Technology-Enhanced Teaching and Learning: Connecting to Psychology and the Social Sciences*, eds D. S. Dunn, J. H. Wilson, J. E. Freeman, and J. R. Stowell (New York, NY: Oxford University Press), 299–316.

Subramanian, K. (2014). "VisiStat: Visualization-driven, interactive statistical analysis," in *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 987–992.

Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th Edn. Boston, MA: Pearson.

Tishkovskaya, S., and Lancaster, G. A. (2012). Statistical education in the 21st century: a review of challenges, teaching innovations and strategies for reform. *J. Stat. Educ.* 20.

Twycross, A., and Shields, L. (2004). Statistics made simple: part 4 choosing the right statistical test. *Paediatr. Nurs.* 16, 24–24. doi: 10.7748/paed2004.10.16.8.24.c944

Unz, D., and Hesse, F. (1999). The use of hypertext for learning. *J. Educ. Comput. Res.* 20, 279–295. doi: 10.2190/FB08-5H94-3R3Y-KQLY

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarter.* 27, 425–478.

Wacharamanotham, C., Subramanian, K., Völkel, S. T., and Borchers, J. (2015). "Statsplorer: Guiding novices in statistical analysis," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 2693–2702. doi: 10.1145/2702123.2702347

Wagner, C., Garner, M., and Kawulich, B. (2011). The state of the art of teaching research methods in the social sciences: Towards a pedagogical culture. *Stud. High. Educ.* 36, 75–88. doi: 10.1080/03075070903452594

Ware, M. E., and Chastain, J, D. (1989). Computer-assisted statistical analysis: a teaching innovation? *Teach. Psychol.* 16, 222–227. doi: 10.1207/s15328023top1604_16

Ware, M. E., and Chastain, J. D. (1991). Developing selection skills in introductory statistics. *Teach. Psychol.* 18, 219–222. doi: 10.1207/s15328023top1804_4

Yan, J., and Lavigne, N. C. (2014). Promoting college students' problem understanding using schema-emphasizing worked examples. *J. Exp. Educ.* 82, 74–102. doi: 10.1080/00220973.2012.745466

Yin, Y. (2012). Using tree diagrams as an assessment tool in statistics education. *Educ. Assess.* 17, 22–49. doi: 10.1080/10627197.2012.697850

Yu, C., Lee, S. J., and Ewing, C. (2014). "Mobile learning: emerging trends, issues, and challenges in teaching and learning," in *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2014*, ed. T. Bastiaens (Chesapeake, VA: Association for the Advancement of Computing in Education), 2126–2136

Thinking Outside the Box: Developing Dynamic Data Visualizations for Psychology with Shiny

David A. Ellis[1]* and Hannah L. Merdian[2]

[1] *Department of Psychology, Lancaster University, Lancaster, UK,* [2] *School of Psychology, University of Lincoln, Lincoln, UK*

The study of human perception has helped psychologists effectively communicate data rich stories by converting numbers into graphical illustrations and data visualization remains a powerful means for psychology to discover, understand, and present results to others. However, despite an exponential rise in computing power, the World Wide Web, and ever more complex data sets, psychologists often limit themselves to static visualizations. While these are often adequate, their application across professional psychology remains limited. This is surprising as it is now possible to build dynamic representations based around simple or complex psychological data sets. Previously, knowledge of HTML, CSS, or Java was essential, but here we develop several interactive visualizations using a simple web application framework that runs under the *R* statistical platform: *Shiny*. *Shiny* can help researchers quickly produce interactive data visualizations that will supplement and support current and future publications. This has clear benefits for researchers, the wider academic community, students, practitioners, and interested members of the public.

Keywords: visualization, knowledge-exchange, research methods, statistics, *R*, Shiny

INTRODUCTION

Psychological data analysis continues to develop with a recent shift in focus from significance testing to the exploration of effect sizes and confidence intervals (Schmidt, 1996; Sainani, 2009). At the same time, psychology and related fields have made meaningful contributions when it comes to developing innovative methods for visualizing and interpreting findings (for a brief history see Friendly, 2008). Historically, the focus has often been to maximize the expressive power of figures, both with regards to conveying the content and structure of the data as well as informing the analysis process (Campitelli and Macbeth, 2014; Marmolejo-Ramos, 2014). This has included a number of computational developments, such as the expansion of boxplots to include information about both distribution and density of the data (Marmolejo-Ramos and Matsunaga, 2009; Marmolejo-Ramos and Tian, 2010) or explorations of different data visualizations for particularly skewed data sets (Ospina et al., 2014).

However, while static graphical illustrations remain perfectly adequate in many instances, these have become problematic as we move toward larger and more complex data sets that evolve over time (Heer and Kandel, 2012). In a critical review concerning the use of data visualizations in scientific papers, Weissgerber et al. (2015) identified a number of limitations and misrepresentations linked to the current practice of using static figures when presenting continuous data from small sample sizes. Static data visualizations are also limited in the quantity and type
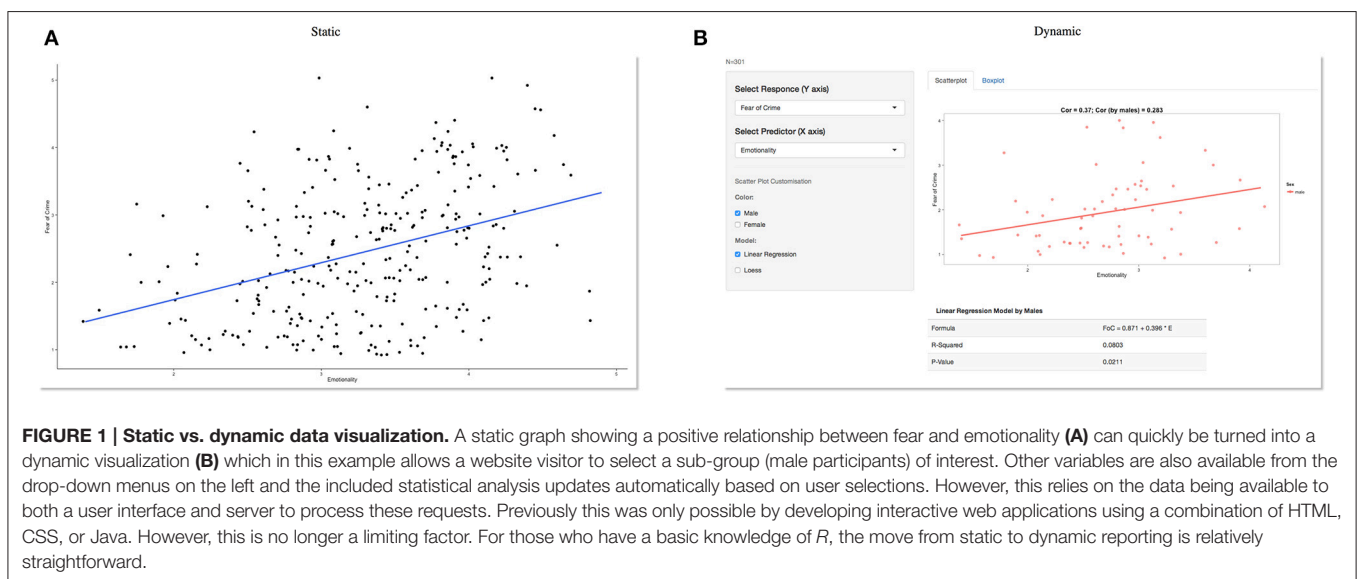
of information that can be presented, which is typically directed toward the analysis conducted. These visualizations in isolation often raise additional questions about the data itself or suggest an alternative analysis. Dynamic representations on the other hand can provide an almost limitless supply of additional information; at a basic level, for example, this would enable a regression model to be re-calculated in real-time for male and female participants separately (**Figure 1**).

Complex applications can also provide online portals for interactive data augmentation and collaboration (Tsuji et al., 2014). However, such transformations rely on the data being available to both a user interface and server to process these requests. Previously this was only possible by developing interactive web applications using a combination of HTML, CSS, or Java, but this is no longer a limiting factor. For those who have a basic knowledge of R, the move from static to dynamic reporting is relatively straightforward (e.g., Xie, 2013).

Dynamic data visualization is likely to have clear advantages when teaching statistical concepts to undergraduate students; for example, Newman and Scholl (2012) pointed toward issues in students' interpretation of bar graphs (a static representation), with Moreau (2015) stating that visual and dynamic data representations may be more appropriate when teaching complex statistical concepts. For example, learning across multiple visual representations has been shown to improve students' understanding (Bodemer et al., 2004). It may also motivate students who were previously of the opinion that becoming statistically literate involves understanding numbers in isolation (Papastergiou, 2009).

Going further, dynamic data visualization can also fulfill the particular research needs of practitioners in the applied sciences including clinical and forensic psychology. One of the core competencies of professional psychologists in practice is to develop an understanding and application of scientific knowledge in evidence-based practice. These competencies should remain closely aligned to the development of methodological skills when evaluating research (e.g., American Psychological Association, 2011; British Psychological Society, 2014). Training is guided by the Scientist-Practitioner Model, postulating that effective psychological services are underpinned by research that is informed by questions arising from clinical practice (Jones and Mehr, 2007). However, there is no professional consensus in terms of the exact nature of the relationship between psychological science and professional practice (Peterson, 2000; Gelso, 2006). In their review of current issues regarding the future development of forensic psychology, Otto and Heilbrun (2002) emphasized practicing forensic psychology in line with the "relevant empirical data" (p. 16) but failed to systematically incorporate the scientific method as a development target for forensic psychologists. Gelso (2006) considers that a low level of research engagement by clinical doctorate graduates (e.g., Barlow, 1981; Peterson et al., 1982; Shinn, 1987) is due to neglect of the research training within the academic environment for professional psychologists, and to a lack of specific research skills required within their professions. Even for those undertaking pure research degrees, Aiken et al. (2008) identified significant gaps in the knowledge of doctoral students with major misunderstandings evident in statistics, measurement, and methodology training, specifically with regards to non-laboratory research, advanced research methods, and innovative methodology and research design. These training gaps constitute a particular disadvantage for clinical and forensic research productivity, where research is often based on single-case studies (e.g., ABA-designs in clinical practice) or small sample sizes (e.g., specific offender or clinical subtypes). Frequently, a large number of variables for each data point are available for a small number of cases that will often not fulfill the assumptions required for traditional linear tests (e.g.,

FIGURE 1 | Static vs. dynamic data visualization. A static graph showing a positive relationship between fear and emotionality **(A)** can quickly be turned into a dynamic visualization **(B)** which in this example allows a website visitor to select a sub-group (male participants) of interest. Other variables are also available from the drop-down menus on the left and the included statistical analysis updates automatically based on user selections. However, this relies on the data being available to both a user interface and server to process these requests. Previously this was only possible by developing interactive web applications using a combination of HTML, CSS, or Java. However, this is no longer a limiting factor. For those who have a basic knowledge of R, the move from static to dynamic reporting is relatively straightforward.

in offender profiling; Canter and Heritage, 1990s). Finally, with the introduction of mobile technology, applied field-research has the capacity to produce very large data sets through the use of mobile applications (e.g., in identifying friendship networks; Eagle et al., 2009; or displaying individual gait patterns; Teknomo and Estuar, 2014). However, both very small and very large data sets provide a challenge for standard linear representations and testing (Rothman, 1990), which we argue can in-part be compensated for with the use of dynamic data visualizations. This would also allow non-experts to repeat (complex) analyses in their own time, after the researcher has provided a summary (Valero-Mora and Ledesma, 2014).

At present, several barriers remain when integrating these methods with psychological research and practice. First, developing suitable applications that can process, analyze and visualize psychological data requires a significant allocation of resources. Second, the lack of concrete examples that directly relate to psychological data mean that current applications are often overlooked. In this tutorial paper, we aim to address both aspects by introducing *Shiny* (http://shiny.rstudio.com/), a data-sharing and visualization platform with low threshold requirements for most psychologists. We then provide several examples centered on a real-life forensic research dataset, which aimed to develop a predictive model for crime-related fear.

INTRODUCING SHINY

Shiny allows for the rapid development of visualizations and statistical applications that can quickly be deployed online. By providing a web application framework for *R* (http://www.r-project.org), this platform allows researchers, practitioners and members of the public to interact with data in real-time and generate custom tables and graphs as required[1].

Shiny applications have two components: a user-interface definition and a server script. These cleverly combine any additional data, scripts, or other resources required to support the application; data can either be uploaded to or retrieved from an online repository. The remainder of this paper will create and develop an interactive visualization using an example data set concerning factors that predict an individual's crime-related fear.

Developing any Shiny app or dynamic data visualization can be split into four steps:

(i) Data preparation
(ii) Creating static content to guide development
(iii) Development and testing
(iv) Deploying an application online

Data Preparation

We recently collected data from around 300 participants which included a variety of variables that might predict an individual's fear of crime (see data.csv in Supplementary Material). While we were particularly interested in personality factors that predict fear, we also collected anxiety and well-being scores along with every participant's age and gender (see **Table 1** for a list of

---

[1]An accompanying website is also available https://sites.google.com/site/psychvisualizations/

TABLE 1 | Information about the included dataset—`data.csv` (Supplementary Material).

Variable	Name in dataset
Participant ID	Participant
Gender*	sex
Age	age
Victim of crime*	victim_crime
Honesty-humility	H
Emotionality	E
Extraversion	X
Agreeableness	A
Conscientiousness	C
Openness to experience	O
State anxiety	SA
Trait anxiety	TA
Happiness	OHQ
Fear of crime	FoC
Fear of crime (2 item version)	Foc2

Copies of this data set can be found in all included code folders (Supplementary Material).
*Categorical variable. Remaining variables are all numeric with higher scores indicating increased levels of each trait.

included variables). We felt that that these findings may be of interest to members of the public and other interested parties (e.g., law enforcement agencies), and wanted to report the results in a dynamic fashion that allow external parties access the data and subsequent results.

The included data set can be loaded into *R* using the read.csv command:

```
data <- read.csv("data.csv", header = T,
sep = ",")
```

An identical dataset crime.csv is included with all example code folders.

Care should be taken by the data provider to only include variables that will be used as part of the final online application; for example, while almost all of our example variables were calculated from an extensive set of standardized measures, including the HEXACO-PI-R measure of personality (Ashton and Lee, 2009), we have not included the raw data for each measure to ensure that the final application will load and update quickly once online.

Creating Static Content to Guide Development

Before creating any *Shiny* application, it is useful to experiment with some simple statistical analysis and static visualization in order to get a feeling for how the data can best be represented within an application. One may conclude that a static visualization (e.g., a single table or series of bar-graphs) is perfectly adequate without any additional development.

Code to install all relevant packages and generate static visualizations in *R* can be found in the `static_graphics` folder. From these examples, we concluded that for our data on crime-related fear, box and scatter plots were ideal when it came to exploring relationships between our variables of interest. Based

on our original predictions, it became evident that specific aspects of personality, such as Emotionality, were likely to be the best predictors of crime-related fear. We also observed that there were a large number of variables and relationships we would like to explore and share with others; however, multiple scatter plots and regression lines would quickly become overwhelming, leading us to develop an application to share our results and data with others.

Development and Testing

We developed a series of examples that progress in complexity. Example 1 makes the simple transition from static to dynamic visualization using a Shiny function. Examples 2 and 3 add advanced customization features using additional graphical and statistical functions.

Example 1

To run the first example, load the *Shiny* library and set your working directory to the folder containing `example1`. This folder includes the data set and two scripts, ui.R and server.R (see below): `library("shiny")`.

The move from static to dynamic visualization only requires a few additional lines of code. The ui.R script loads and labels the variables from the dataset. Here, we aimed to demonstrate how different personality factors might predict an individual's fear of crime, so these are labeled as responses and predictors accordingly. The second part of this script creates a simple *Shiny* page; various placeholders allow users to interact with the data. Finally, a command to print graphical output is placed at the end of this loop.

Moving to the server.R script, variable names defined within ui.R are replicated here. These variable names act as a link between both scripts. An *IF* function provides additional user interaction by differentiating between participants' gender. For example, if male, female or both genders are selected, then the chart will color each data point accordingly. If no participant gender is selected, then a standard plot is created that includes data from both male and female participants.

To run this example, simply type: `runApp('example1')` into the console. A scatter plot should now appear in a new window with a variety of options on the left ("Select Response," "Select Predictor"). By experimenting with different predictors, the scatter plot will update accordingly; this process will assist the development of future predictions regarding what individual differences are more predictive of crime-related fear than others.

Examples 2 and 3[2]

Examples 2 and 3 are developed directly from Example 1. Marked-up code is available in the Supplementary Material, `example2` and `example3`. These can be run in an identical fashion to `example1`. Example 2 adds boxplots and statistical output, which again relies on standard graphical and mathematical functions in *R*. This version also allows the user to build linear regression models after choosing any predictor and response variable (e.g., the predictive value of

Honest-Humility); statistical output is presented underneath the scatter plot, providing information relating to effect sizes and statistical significance. Box plots can be used to directly compare the distribution of scores on these variables, or to compare levels of crime-related fear between men and women directly. Example 3 (**Figure 2**) adds two additional functions, which handle a variety of potential visualization options. This provides separate regression outputs for male and female participants and/or those who have previously been a victim of crime.

Deploying an Application Online[3]

There are several ways to deploy a *Shiny* application online; however, the fastest route is to create a Shiny account (http://www.shinyapps.io/) and install the *devtools* package by running the following code in your R console: `install.packages('devtools')`.

Finally, the *rsconnect* package is also required and can be installed by running the following code in your R console: `devtools::install_github('rstudio/rsconnect)`. Load this library: `library("rsconnect")`. Once a shinyapps.io account has been created online and authorized, any of the included examples can quickly be deployed straight from the R console: `deployApp("example1")`. However, it is also possible to host your own private *Shiny* server[4].

Deployment of the application will allow anyone with an internet connection to engage with the data directly. However, the entire dataset could also be made available from the application itself with some additional development.

DISCUSSION

The last two decades have witnessed marked changes to the use and implementation of data visualizations. While research has often focused on the enhancement of existing static visualization tools, such as violin plots to express both density and distribution of data (Marmolejo-Ramos and Matsunaga, 2009), these remain limited due to their static nature. Specifically, static visualizations become exponentially more difficult to understand as the complexity of the content they aim to display increases (e.g., Teknomo and Estuar, 2014).

Such data-rich representations are likely to be helpful when teaching statistical concepts however, little research exists on its effectiveness within an educational context (Valero-Mora and Ledesma, 2014). While an expert user may believe they have created something practical and aesthetically pleasing, much of the literature surrounding human-computer interaction repeatedly demonstrates how a seemingly straightforward system that an expert considers "easy" to operate often poses significant challenges to new users (Norman, 2013). Future research is required in order to fully understand the effect interactive visualizations could have on a student's understanding of complex statistical concepts.

Dynamic visualizations remain a promising alternative to display and communicate complex data sets in an accessible

---

[2]Example 3 can be viewed online https://psychology.shinyapps.io/example3

[3]Additional instructions are available http://shiny.rstudio.com/articles/shinyapps.html

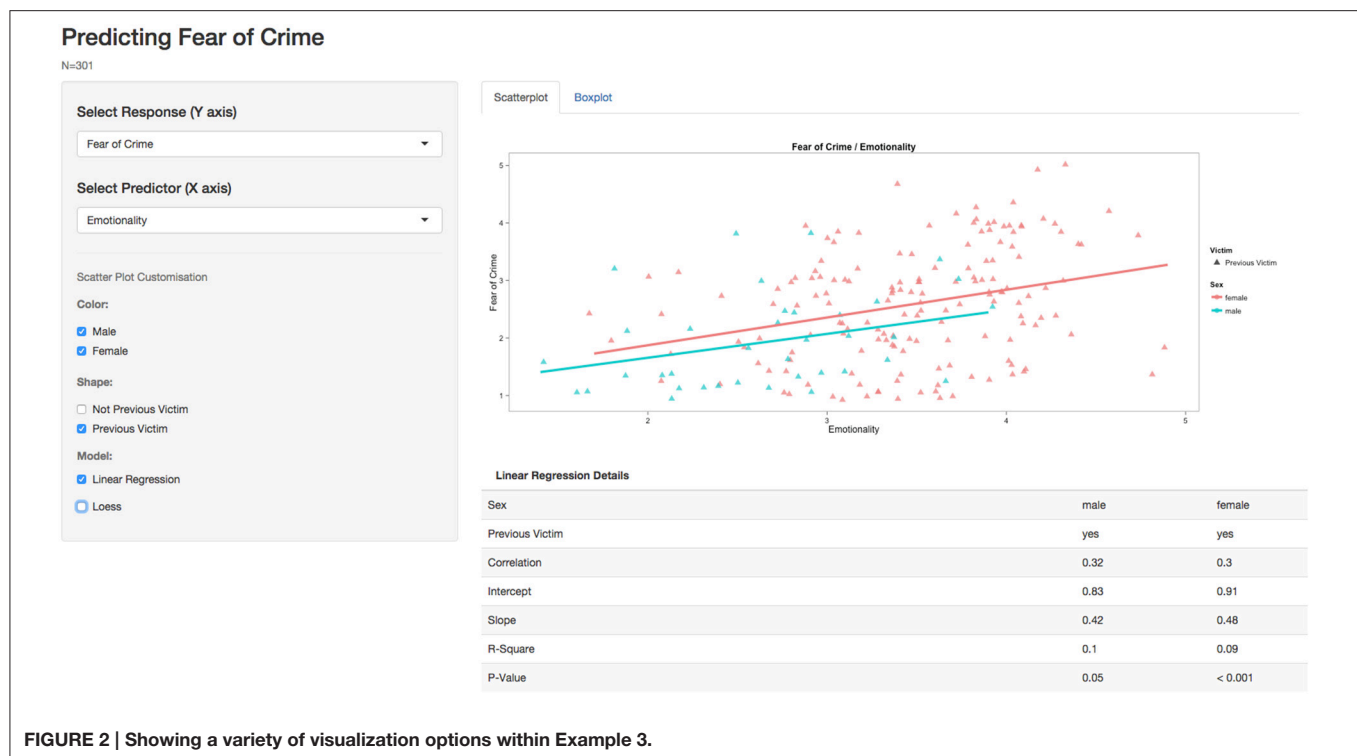[4]http://www.rstudio.com/products/shiny/download-server/

FIGURE 2 | Showing a variety of visualization options within Example 3.

manner for expert and non-expert audiences (Valero-Mora and Ledesma, 2014). The above worked examples demonstrate the straightforward and flexible nature of dynamic visualization tools such as *Shiny*, using a real-life example from forensic psychology. This move toward a more dynamic graphical endeavor speaks positively toward cumulative approaches to data aggregation (Braver et al., 2014), but it can also provide non-experts with access to simple and complex statistical analysis using a point-and-click interface. For example, through exploration of our fear of crime data set, it should quickly become apparent that while some aspects of personality do correlate with fear of crime, the results are not clear-cut when considering men and women in isolation and this may generate new hypotheses concerning gender differences and how a fear of crime is likely to be mediated by other variables.

While a basic knowledge of *R* is essential, dynamic visualizations can make a technically proficient user more productive, while also empowering students and practitioners with limited programming skills. For example, an additional *Shiny* application could automatically plot an individual's progress throughout a forensic or clinical intervention. Relationships between variables of improvement alongside pre and post scores across a several measures could also be displayed in real-time with results accessible to clinicians and clients. Dynamic data visualizations may therefore be the next step toward bridging the gap between scientists and practitioners.

The benefits to psychology are not simply limited to improved understanding and dissemination, but also feed into issues of

replication. For example, the ability to compare multiple or pairs of replications side by side is now possible by providing suitable user interfaces. Tsuji et al. (2014), for example, have recently developed the concept of community-augmented meta-analysis (CAMA), which involves a combination of meta-analysis and an open repository (e.g., PsychFileDrawer.org; Spellman, 2012). These alone can improve research practices by ensuring that past research is integrated into current work. Using the intervention example from above, one can envision a further application that plots the progress of individual clients over several years, providing information on treatment change, outliers, and group trends over time.

In other areas of psychological research, much of this data already exists and the availability of data on open access repositories (e.g., such as *Dryad* or *Figshare*) makes data deposition in the first instance more straightforward. However, the advantages of open-access databases brings with it problems of navigation, organization and understanding. If these new developments are to reach their full potential and remain relevant to all psychologists, they still require a user-friendly interface that allows for rapid re-analysis and visualization. Of course, dynamic or interactive data visualizations are only going to become standard practice if psychologists use these methods on a regular basis. Researchers themselves will govern the speed of this development; journals may start to support this additional interactivity within publications. We hope that in addition to providing open access to data, psychologists will also popularize the shift toward dynamic visualizations in basic and applied research.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01782

REFERENCES

Aiken, L. S., West, S. G., and Milsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology replication and extension of aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *Am. Psychol.* 63, 32–50. doi: 10.1037/0003-066X.63.1.32

American Psychological Association (2011). *Revised Competency Benchmarks for Professional Psychology.* Available online at: http://www.apa.org/ed/graduate/competency.aspx

Ashton, M. C., and Lee, K. (2009). The HEXACO-60: a short measure of the major dimensions of personality. *J. Pers. Assess.* 91, 340–345. doi: 10.1080/00223890902935878

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: current issues, new directions. *J. Consult. Clin. Psychol.*, 49, 147–155. doi: 10.1037/0022-006X.49.2.147

Bodemer, D., Ploetzner, R., Feuerlein, I., and Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learn. Instruct.* 14, 325–341. doi: 10.1016/j.learninstruc.2004.06.006

Braver, S. L., Thoemmes, F. J., and Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.* 9, 333–342. doi: 10.1177/1745691614529796

British Psychological Society (2014). *Standards for Doctoral Programmes in Clinical Psychology.* Available online at: http://www.bps.org.uk/system/files/Public%20files/PaCT/dclinpsy_standards_approved_may_2014.pdf

Campitelli, G., and Macbeth, G. (2014). Hierarchical graphical Bayesian models in psychology. *Rev. Colomb. Estadística* 37, 319–339. doi: 10.15446/rce.v37n2spe.47940

Canter, D. V., and Heritage, R. (1990). A multivariate model of sexual offences behaviour: developments in 'offender profiling'. *Int. J. Forensic Psychiatry* 1, 185–212. doi: 10.1080/09585189008408469

Eagle, N., Pentland, A. S., and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci.* 106, 15274–15278. doi: 10.1073/pnas.0900282106

Friendly, M. (2008). The golden age of statistical graphics. *Stat. Sci.* 23, 502–535. doi: 10.1214/08-sts268

Gelso, C. J. (2006). On the making of a scientist–practitioner: a theory of research training in professional psychology. *Train. Educ. Prof. Psychol.* S(1), 3–16. doi: 10.1037/1931-3918.S.1.3

Heer, J., and Kandel, S. (2012). Interactive analysis of big data. XRDS: Crossroads. *ACM Mag. Stud.* 19, 50–54. doi: 10.1145/2331042.2331058

Jones, J. L., and Mehr, S. L. (2007). Foundations and assumptions of the scientist-practitioner model. *Am. Behav. Sci.* 50, 766–771. doi: 10.1177/0002764206296454

Marmolejo-Ramos, F. (2014). Editorial. Current topics in statistical graphics [editorial]. *Rev. Colomb. Estadística* 37, 1–4. doi: 10.15446/rce.v37n2spe.48058

Marmolejo-Ramos, F., and Matsunaga, M. (2009). Getting the most from your curves: exploring and reporting data using informative graphical techniques. *Tutor. Quant. Methods Psychol.* 5, 40–50.

Marmolejo-Ramos, F., and Tian, T. S. (2010). The shifting boxplot. A boxplot based on essential summary statistics around the mean. *Int. J. Psychol. Res.* 3, 37–45.

Moreau, D. (2015). When seeing is learning: dynamic and interactive visualizations to teach statistical concepts. *Front. Psychol.* 6:342. doi: 10.3389/fpsyg.2015.00342

Newman, G. E., and Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: the within-the-bar bias. *Psychon. Bull. Rev.* 19, 601–607. doi: 10.3758/s13423-012-0247-5

Norman, D. A. (2013). *The Design of Everyday Things, Revised and Expanded Edition.* New York, NY: Basic Books.

Ospina, R., Larangeiras, A. M., and Frery, A. C. (2014). Visualization of skewed data: a tool in R. *Rev. Colomb. Estadística* 37, 399–417. doi: 10.15446/rce.v37n2spe.47945

Otto, R. K., and Heilbrun, K. (2002). The practice of forensic psychology: a look toward the future in light of the past. *Am. Psychol.* 57, 5–18. doi: 10.1037/0003-066X.57.1.5

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Comput. Educ.* 52, 1–12. doi: 10.1016/j.compedu.2008.06.004

Peterson, D. R. (2000). Scientist-practitioner or scientific practitioner? *Am. Psychol.* 55, 252–253 doi: 10.1037/0003-066X.55.2.252

Peterson, D. R., Eaton, M., Levine, R., and Snepp, F. (1982). Career experiences of doctors of psychology. *Prof. Psychol.* 13, 274–285.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* 1, 43–46. doi: 10.1097/00001648-199001000-00010

Sainani, K. L. (2009). The problem of multiple testing. *Phys. Med. Rehabil.* 1, 1098–1103. doi: 10.1016/j.pmrj.2009.10.004

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for the training of researchers. *Psychol. Methods* 1, 115–129.

Shinn, M. R. (1987). Research by practicing school psychologists: the need for fuel for the lamp. *Prof. School Psychol.* 2, 235–243. doi: 10.1037/h0090549

Spellman, B. A. (2012). Introduction to the special section on research practices. *Perspect. Psychol. Sci.* 7, 655–656. doi: 10.1177/1745691612465075

Teknomo, K., and Estuar, M. R. (2014). Visualizing gait patterns of able bodied individuals and transtibial amputees with the use of accelerometry in smart phones. *Rev. Colomb. Estadística* 37, 471–488. doi: 10.15446/rce.v37n2spe.47951

Tsuji, S., Bergmann, C., and Cristia, A. (2014). Community-augmented meta-analyses toward cumulative data assessment. *Perspect. Psychol. Sci.* 9, 661–665. doi: 10.1177/1745691614552498

Valero-Mora, P., and Ledesma, R. (2014). Dynamic-interactive graphics for statistics (26 years later). *Rev. Colomb. Estadística* 37, 247–260. doi: 10.15446/rce.v37n2spe.47932

Weissgerber, T. L., Milic, N. M., Winham, S. J., and Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol.* 13:e1002128. doi: 10.1371/journal.pbio.1002128

Xie, Y. (2013). animation: An R package for creating animations and demonstrating statistical methods. *J. Stat. Softw.* 53, 1–27. doi: 10.18637/jss.v053.i01

When seeing is learning: dynamic and interactive visualizations to teach statistical concepts

David Moreau *

Centre for Brain Research, University of Auckland, Auckland, New Zealand

When Seeing is Learning: Dynamic and Interactive Visualizations to Teach Statistical Concepts
"So, how would I use statistics in psychological research? First of all, descriptively."

– Jacob Cohen (1990), p. 1310.

Jacob Cohen, one of the greatest statisticians of the twentieth century, reflected upon a problem many students of introductory statistics courses can relate to: the benefit of clear visual representations to understand statistical properties. Naturally, Cohen was not the first to touch upon this idea; in fact, he was echoing another exceptional statistician of the last century, John Tukey, who had emphasized the necessity of depicting data visually (Tukey, 1977). Decades later, this idea is still very contemporary—technologies have evolved, computing power has soared, but the perplexing nature of data analysis remains, especially for young scientists (Watts, 1991; Garfield and Ben-Zvi, 2007).

What advances in computing allow, however, is a fresh approach to circumvent the problem. Modern technologies offer an impressive panel of tools and support to teach statistics, in and outside the classroom. It is now possible for students to dynamically interact with their data, in order to understand the relative contribution of individual data points, variables, or parameters. For example, they can fairly easily introduce one data point at a time in an analysis and monitor how each observation refines the underlying model. Or observe how changes in one parameter result in differences in power calculations (**Figure 1**). Students can also access fairly abstract concepts such as randomness or sampling in a glimpse.

Moreover, these advances come at a time when statistics is becoming a seductive and appealing field to students and the general public. In the New York Times and the Wall Street Journal, at Google and Facebook, the message is clear: statistics *is* sexy. Equally revealing, Hans Rosling, in a riveting and broadly-praised talk, advises not to be ashamed to say you are a statistician at a dinner party (Rosling, 2010). We have come a long way.

Yet how can we allow more students to grasp the critical concepts required to become statistically literate and apply data analysis methods adequately in current and future research programs? Many suggestions have been made, and this special research topic provides additional fascinating leads. In my opinion, one of the answers lies in the diversification of teaching contents and media. Decades of research have shown that individuals differ greatly in their ability to generalize, maintain and manipulate mental images. For example, while some need to constantly switch their focus of attention between the target figure and the possible answers in a mental rotation task, others can identify the correct answer in a blink, performing with remarkable accuracy in an effortless manner (Hegarty and Waller, 2005). It is therefore no coincidence one of the most eminent pioneers in the study of mental imagery has provided guidance on how to create effective visual presentations–he knows firsthand the tremendous variability among individuals when it comes to visualization, and the power of clear visual depictions to convey information accessible

FIGURE 1 | Example of power analysis for a two-tailed symmetrical test. With R packages such as shiny, students can modify any of the four quantities (α, β-1, ES, and N) and understand visually how their manipulation affects one another. Sliders can be added to the graph to facilitate manipulation (omitted here for clarity purposes).
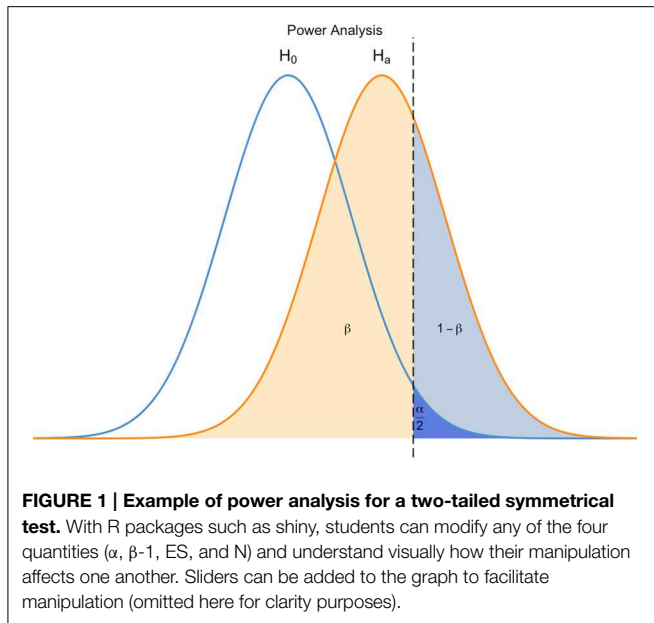
to everyone (Kosslyn, 2007). Interindividual variability in spatial ability also has important consequences in the way statistics should be taught in the classroom. An instructor cannot expect all students to extract the same information from a graphical depiction, or to be equally comfortable with complex representations of data. Because of these discrepancies, any effort to make visual content more accessible should be encouraged. Via dynamic and interactive graphics, today's technology allows students to visualize externally what they have difficulty representing mentally.

Dynamic and interactive visualizations also allow learning by active exploration—students can engage with their data, rather than try to understand them passively. Active engagement results in improved learning and verbal understanding (Bodemer et al., 2004) and is especially important since direct interaction with visual content facilitates the involvement of the motor system (Wraga et al., 2003), a component often present in highly-visual individuals and crucial to achieve deeper levels of understanding. Data exploration can then become a multi-sensory experience, setting the stage for profound and effective learning. Interactive visualizations are also more enjoyable, and in that respect possess prime value to motivate reluctant students (Papastergiou, 2009).

These advantages come at a cost. They require additional work beforehand on the teacher's part, to integrate components that are essential for valid, powerful pedagogical content. Graphics should include relevant content, but no more–too often, visual depictions contain more information than one can possibly process, resulting in cognitive overload and additional effort

to consider information by chunks at a time (Lowe, 2003). For example, content from three-dimensional depictions could often be presented with separate two-dimensional plots, easier to comprehend. Just because a software offer the possibility of advanced graphics does not mean it is always the most appropriate choice. Simple, well-thought representations are often more effective. Directly related to this idea, changes in appearance should reflect new information, rather than diversification for esthetic purposes. When it comes to representing data, people expect changes in features to carry information (Kosslyn, 2007). Finally, visualizations need to be accessible to anyone, which means working toward user-friendly, non-discriminatory content. For example, graphics should avoid hues indiscernible by color-blind individuals. Wide accessibility also means opting for free software whenever possible, to guarantee sustained and undisrupted access. In this regard, graphical representations are made extremely easy by the growing popularity of statistical software such as R (R Core Team, 2014). Besides being free and open-source, allowing students to explore and play with their data from anywhere, R makes dynamic, interactive visualizations simple, with packages such as shiny, googleVis, and rCharts. R also lets creative instructors build *ad hoc* scripts and packages for their teaching purposes, thus enabling any type of data visualization.

Another definite asset of R and of the new generation of software pertains to the simplicity of simulating data. It is often crucial for students to make sense of a concept visually before transitioning to more sophisticated mathematical models. Despite the obvious advantages in understanding data, visual inputs can lack the general aspect of equations–they reflect a particular pattern of data at a specific time, yet to fully grasp most statistical concepts, students need to build an internal representation general enough to be useful in a wide range of situations. Simulations bridge that gap, as they help derive rules from the dynamic exploration of visual outputs. Other possibilities exist for implementing simulations, such as Java applets, but R has the remarkable advantage of including almost all a student needs in one single software.

In closing, it is a great time to be studying statistics—between online resources, MOOCs, free software and the appeal of data science to the academic and professional worlds, everything seems well-aligned to reward a curriculum emphasizing statistics. It is equally exhilarating to be teaching statistics—new technologies offer suitable tools to personalize content and reach more students. If any, the downside may be that there will soon be no excuse for imprecise knowledge or inaccurate applications of statistical techniques. As teachers and instructors embrace this technological revolution, a statistics-literate generation of psychologists will emerge, less prone to misuse statistics, and more likely to advance scientific knowledge while conveying clear, understandable messages to general audiences.

References

Bodemer, D., Ploetzner, R., Feuerlein, I., and Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learn. Instr.* 14, 325–341. doi: 10.1016/j.learninstruc.2004.06.006

Cohen, J. (1990). Things I have learned (so far). *Am. Psychol.* 45, 1304–1312. doi: 10.1037/0003-066X.45.12.1304

Garfield, J., and Ben-Zvi, D. (2007). How students learn statistics revisited: a current review of research on teaching and learning statistics. *Int. Stat. Rev.* 75, 372–396. doi: 10.1111/j.1751-5823.2007.00029.x

Hegarty, M., and Waller, D. A. (2005). "Individual differences in spatial abilities," in *The Cambridge Handbook of Visuospatial Thinking,* eds P. Shah and A. Miyake (New York, NY: Cambridge University Press), 121–169. doi: 10.1017/CBO9780511610 448.005

Kosslyn, S. M. (2007). *Clear and to the Point: 8 Psychological Principles for Compelling PowerPoint Presentations.* New York, NY: Oxford University Press.

Lowe, R. K. (2003). Animation and learning: selective processing of information in dynamic graphics. *Learn. Instr.* 13, 157–176. doi: 10.1016/S0959-4752(02)00018-X

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Comput. Educ.* 52, 1–12. doi: 10.1016/j.compedu.2008. 06.004

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: http:// www.r-project.org

Rosling, H. (2010). *The Joy of Stats.* BBC Four Documentary.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Watts, D. G. (1991). Why is introductory statistics difficult to learn? And what can we do to make it easier? *Am. Stat.* 45, 290–291.

Wraga, M., Thompson, W. L., Alpert, N. M., and Kosslyn, S. M. (2003). Implicit transfer of motor strategies in mental rotation. *Brain Cogn.* 52, 135–143. doi: 10.1016/S0278-2626(03)00033-2

Recommended Online Resources

https://developers.google.com/chart/interactive/docs/gallery
https://www.geogebra.org/
http://rcharts.io/
http://rpsychologist.com/
http://shinyapps.org/
http://shiny.rstudio.com/
http://showmeshiny.com/
http://swirlstats.com/

The meaning of significance in data testing

Jose D. Perezgonzalez *

Business School, Massey University, Palmerston North, New Zealand

Recent developments in psychology (e.g., Nuzzo, 2014; Trafimow, 2014; Woolston, 2015a) are showing apparently reasonable but inherently flawed positions against data testing techniques (often called hypothesis testing techniques, even when they do not test hypotheses but assume them true for testing purposes). These positions are such as banning testing explicitly and most inferential statistics implicitly (Trafimow and Marks, 2015, for *Basic and Applied Social Psychology*—but see Woolston, 2015a, expanded in http://www.nature.com/news/psychology-journal-bans-p-values-1.17001), recommending substituting confidence intervals for null hypothesis significant testing (NHST) explicitly and for all other data testing implicitly (Cumming, 2014, for *Psychological Science*—but see Perezgonzalez, 2015a; Savalei and Dunn, 2015), and recommending research preregistration as a solution to the low publication of non-significant results (e.g., Woolston, 2015b). In reading Woolston's articles, readers' comments to such articles, and the related literature, it appears that philosophical misinterpretations of old, already discussed by, for example, Meehl (1997), Nickerson (2000), Kline (2004), and Goodman (2008), are not getting through and still need to be re-addressed today. I believe that a chief source of misinterpretations is the current NHST framework, an incompatible mishmash between the testing theories of Fisher and of Neyman-Pearson (Gigerenzer, 2004). The resulting misinterpretations have both a statistical and a theoretical background. Statistical misinterpretations of *p*-values have been addressed elsewhere (Perezgonzalez, 2015c), thus I reserve this article for resolving theoretical misinterpretations regarding statistical significance.

The main confusions regarding statistical significance can be summarized in the following seven points (e.g. Kline, 2004): (1) significance implies an important, real effect size; (2) no significance implies a trivial effect size; (3) significance disproves the tested hypothesis; (4) significance proves the alternative hypothesis; (5) significance exonerates the methodology used; (6) no significance is explainable by bad methodology; and (7) no significance in a follow up study means a replication failure. These seven points can be discussed according to two concerns: the meaning of significance itself, and the meaning, or role, of testing.

In this article I will avoid NHST and, instead, refer to either Fisher's or Neyman-Pearson's approaches, when appropriate. I will also avoid their conceptual mix-up by using different concepts, those which seem most coherent under each approach. Thus, Fisher's seeks significant results, tests data on a null hypothesis (H_0) and uses levels of significance (sig) to ascertain the probability of the data under H_0 (**Figure 1A**). Neyman-Pearson's seeks to make a decision, tests data on a main hypothesis (H_M) and decides in favor of an alternative hypothesis (H_A) according to a cut-off calculated a priori based on sample size (N), Type I error probability (α), effect size (MES) and power ($1-\beta$), the latter two provided by H_A (**Figure 1B**).
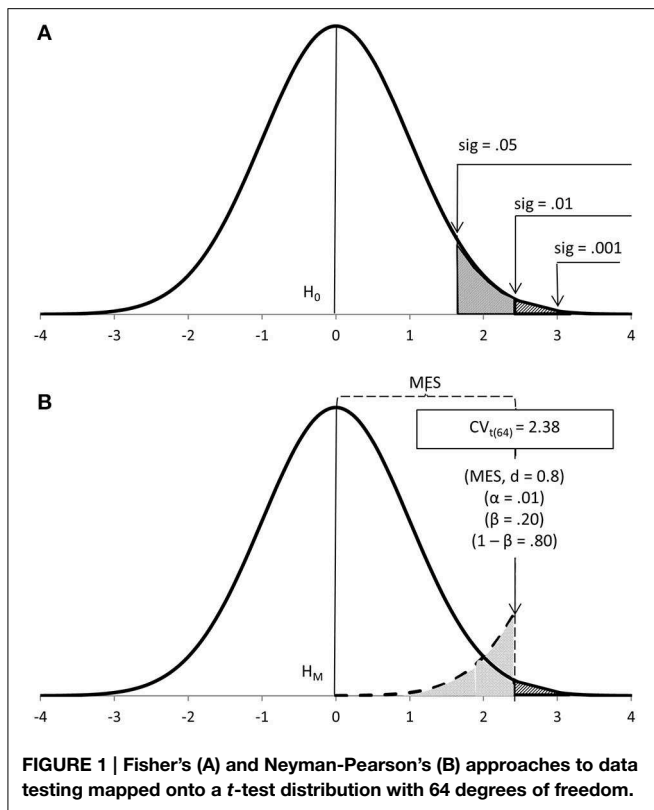
FIGURE 1 | Fisher's (A) and Neyman-Pearson's (B) approaches to data testing mapped onto a *t*-test distribution with 64 degrees of freedom.

The Meaning of Significance

Misinterpretations (1) and (2) are due to confusing statistical significance, theoretical or practical significance, and effect sizes. The latter are a property of populations, may vary from large to small, can be put into a number, and can be calculated with the appropriate formula (Cohen, 1988). Practical significance is a subjective assessment of the importance of such effect (they can be considered as two sides of the same coin).

As per statistical significance, because testing is done on samples, it is the equivalent population effects in the corresponding sampling distributions which are of relevance, to be found in the tail (or tails) of such distributions. When using these techniques, therefore, important effects become extreme results—i.e., results with low *p*-values—under the tested hypotheses.

To help with inferences, a cut-off is used to partition the corresponding sampling distribution between extreme and not-extreme-enough results. This is, of course, a pragmatic choice, but it implies that the meaning of significance ultimately depends on where such cut-off falls. This cut-off also partitions the effect size between important and unimportant effects for testing purposes.

Under Neyman-Pearson's approach (e.g., 1933; **Figure 1B**), the mathematically-set cut-off partitions the effect size a priori (the minimum effect size, MES, is the value of the population effect size at such point; Perezgonzalez, 2015b). Statistical significance, thus, has no inherent meaning under this approach other than to identify extreme results beyond the set cut-off.

Because the sample size is controlled, mainly to ensure power, such extreme results are not only more probable under H_A but also reflect important population effects.

Under Fisher's approach (e.g., 1954, 1960; **Figure 1A**), either experience-driven or conventional cut-offs help flag noteworthy results—these are properly significant, as in "notable," "worthy of attention"—whose primary value is in their role as evidence for rejecting H_0. Because there is no inherent control of sample size, a large sample may be used if it leads to the rejection of H_0 more readily—thus, a significant result is technically important. Whether it is really important, however, we cannot know (we ought to wait and calculate the effect size a posteriori), but we may assume an unknown MES with boundaries at the appropriate level of significance. Posterior calculations normally shows that when the sample is small, significant results reflect large effect sizes; as the sample grows larger, the resulting effect sizes may shrink into triviality.

Curiously, then, misinterpretations (1) and (2) are only possible under Fisher's approach depending ultimately on the size of the sample used. With small samples—which is the paradigm that Fisher developed—significant results normally do reflect important effects—thus, (1) is typically not a misinterpretation—however some of the non-significant results may also reflect sizeable effects—thus, misinterpretation (2) is still possible. The opposite occurs with larger samples: Effect sizes may be of any size, including negligible ones, and still turn out statistically significant—thus, misinterpretation (1) is plausible—while non-significant results will often be negligible—thus, misinterpretation (2) is not so, but a correct interpretation.

Under Neyman-Pearson's approach, on the other hand, effect sizes are those of populations, known (or fixed) before conducting the research. These effect sizes can, of course, be set differently by different researchers, yet such decision has a technical consequence on the test thereof: It makes a posteriori interpretations of effect sizes meaningless. Thus, an extreme result—accepted under H_A—is always important because the researcher decided so when setting the test; a not-so-extreme result—accepted under H_M—is always trivial for similar reasons; therefore, as far as any particular test result is concerned, (1) and (2) cannot be considered misinterpretations proper under this approach.

The Meaning of Data Testing

The remaning misinterpretations have to do with confusing research substance and data testing technicality. Meehl (1997) provided clear admonition about the substantive aspects of theory appraisal. He set down a conceptual formula for correlating a set of observations with a theory and related components. His formula includes not only the theory under test—from which the statistical hypothesis supposedly flaws—but also auxiliary theories, the everything-else-being-equal assumption (*ceteris paribus*), and reporting quality—all of which address misinterpretations (3) and (4)—as well as methodological quality—which addresses misinterpretations (5) and (6). Thus, the observation of a significant or extreme result is, at most, able to falsify only the conjunction of elements in the

formula instead of the theory under test—i.e., either the theory is false, or the auxiliary theories are false, or the *ceteris paribus* clause is false, or the particulars reported are wrong, or the methodology is flawed. Furthermore, Meehl argues, following the Popperian dictum a theory cannot be proved, so a non-significant or not-extreme-enough result cannot be used for such purpose, either.

Meehl may have slipped on the technicality of testing, though, still confusing a substantive hypothesis—albeit a very specific one—with a statistical hypothesis. Technically speaking, a statistical hypothesis (H_0, H_M, H_A) provides the appropriate frequency distribution for testing research data and, thus, needs to be (technically) true. Therefore, these hypotheses cannot be either proved or disproved—i.e., disproving a statistical hypothesis invalidates both the test and the results used to disprove it! From this follows that the gap between the statistical hypothesis and the related substantive hypothesis that supposedly flaws from the theory under appraisal cannot be closed statistically but only epistemologically (Perezgonzalez, 2015b).

Therefore, misinterpretations (3) and (4) have conflating technical and substantive causes. Meehl's (1997) formula resolves the substantive aspect, while a technical argument can also be advanced as a solution: Statistical hypotheses need to be (assumed) true and, thus, can be neither proved nor disproved by the research data.

As for misinterpretations (5) and (6), about methodology, these too are resolved by Meehl's formula. Methodological quality is a necessary element for theory appraisal, yet also an independent element in the formula; thus, we may observe a particular research result independently of the quality of the methods used. This is something which is reasonable and may need no further discussion, yet it is also something which tends to appear divorced from the research process in psychological reporting. Indeed, psychological articles tend to address research limitations only at the end, in the discussion and conclusion section (see, for example, American Psychological Association's style recommendations, 2010), something which reads more as an act of contrition than as reassurance that those limitations have been taken into account in the research.

Finally, a technical point can also be advanced for resolving the replication misinterpretation (7). Depending on the approach used, replication necessitates either of a cumulative meta-analysis (Fisher's approach; Braver et al., 2014) or of a count of the number of replications carried out (Neyman-Pearson's approach; Perezgonzalez, 2015a,d). A single replication may suffix the former, yet it is the significance of the meta-analysis, not of the individual studies, that counts. As for the latter, one would expect a minimum number of replications (i.e., four) in order to ascertain the power of the study (i.e., a minimum of four successful studies out of five for ascertaining 80% power); a single replication is, thus, not enough. Therefore, the significance or extremeness of a single replication cannot be considered enough ground for either supporting or contradicting a previous study.

Corollary

Late developments in the editorial policies for the journals *Basic and Applied Social Psychology,* and *Psychological Science* aim to improve the quality of the papers submitted for publication (similar attempts have already been attempted in the past—e.g., Loftus, 1993; Kendall, 1997—with rather limited success—e.g., Finch et al., 2004; Fidler et al., 2005). They do so by banning or strongly discouraging the use of inferential tools, more specifically data testing procedures. There are important theoretical and philosophical reasons for supporting the banning of NHST (e.g., Nickerson, 2000), but these do not necessarily extend to either Fisher's or Neyman-Pearson's procedures or to the remaining of the inferential toolbox. The main problem seems to lie with misinterpretations borne out of NHST and the way statistics is taught. *P*-values are often misinterpreted as providing information they do not—something that may be resolved by simply substituting frequency-based heuristics for the probabilistic heuristics currently used (e.g., Perezgonzalez, 2015c). On the other hand, statistical significance is often misinterpreted as practical importance. Substantive arguments about theory appraisal can resolve some of the misinterpretations, although this requires some reading about epistemology (e.g., Meehl, 1997). Furthermore, technical arguments can also be advanced to resolve other misinterpretations. Many of these confusions could be easily captured and prevented at pedagogical levels, thus highlighting the important role of doing so when teaching statistics. The risk of not doing so is transferred forward to the rest of psychology, which may suffer when misunderstood testing procedures and other inferential tools are discouraged or banned outright for the purpose of, paradoxically, improving psychological science.

References

American Psychological Association. (2010). *Publication Manual of the American Psychological Association, 6th Edn.* Washington, DC: APA.

Braver, S. L., Thoemmes, F. J., and Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.* 9, 333–342. doi: 10.1177/1745691614529796

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* New York, NY: Psychology Press.

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., et al. (2005). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. *J. Consult. Clin. Psychol.* 73, 136–143. doi: 10.1037/0022-006X.73.1.136

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., et al. (2004). Reform of statistical inference in psychology: the case of memory & cognition. *Behav. Res. Methods Instrum. Comput.* 36, 312–324. doi: 10.3758/BF03195577

Fisher, R. A. (1954). *Statistical Methods for Research Workers, 12th Edn.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1960). *The Design of Experiments, 7th Edn.* Edinburgh: Oliver and Boyd.

Gigerenzer, G. (2004). Mindless statistics. *J. Soc. Econ.* 33, 587–606. doi: 10.1016/j.socec.2004.09.033

Goodman, S. N. (2008). A dirty dozen: twelve P-value misconceptions. *Semin. Hematol.* 130, 995–1004. doi: 10.1053/j.seminhematol.2008.04.003

Kendall, P. C. (1997). Editorial. *J. Consult. Clin. Psychol.* 65, 3–5. doi: 10.1037/h0092570

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: APA.

Loftus, G. R. (1993). Editorial comment. *Mem. Cogn.* 21, 1–3. doi: 10.3758/bf03211158

Meehl, P. E. (1997). "The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions," in *What If There Were No Significance Tests?* eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah, NJ: Erlbaum), 393–425.

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. T. R. Soc. A* 231, 289–337. doi: 10.1098/rsta.1933.0009

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989x.5.2.241

Nuzzo, R. (2014). Statistical errors. P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature* 506, 150–152. doi: 10.1038/506150a

Perezgonzalez, J. D. (2015a). Confidence intervals and tests are two sides of the same research question. *Front. Psychol.* 6:34. doi: 10.3389/fpsyg.2015.00034

Perezgonzalez, J. D. (2015b). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* 6:223. doi: 10.3389/fpsyg.2015.00223

Perezgonzalez, J. D. (2015c). P-values as percentiles. *Front. Psychol.* 6:341. doi: 10.3389/fpsyg.2015.00341

Perezgonzalez, J. D. (2015d). Commentary: continuously cumulating meta-analysis and replicability. *Front. Psychol.* 6:565. doi: 10.3389/fpsyg.2015.00565

Savalei, V., and Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Front. Psychol.* 6:245. doi: 10.3389/fpsyg.2015.00245

Trafimow, D. (2014). Editorial. *Basic Appl. Soc. Psych.* 36, 1–2. doi: 10.1080/01973533.2014.865505

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psych.* 37, 1–2. doi: 10.1080/01973533.2015.1012991

Woolston, C. (2015a). Psychology journal bans P values. *Nature* 519:9. doi: 10.1038/519009f

Woolston, C. (2015b). Online debate erupts to ask: is science broken? *Nature* 519:393. doi: 10.1038/519393f

P-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations"

Jose D. Perezgonzalez

Business School, Massey University, Palmerston North, New Zealand

A commentary on:

Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations
by Schneider, J. W. (2015). Scientometrics 102, 411–432. doi: 10.1007/s11192-014-1251-5

Schneider's (2015) article is contemporary work addressing the shortcomings of null hypothesis significance testing (NHST). It summarizes previous work on the topic and provides original examples illustrating NHST-induced confusions in scientometrics. Among the confusions cited are those associated with the interpretation of *p*-values, old misinterpretations already investigated by Oakes (1986), Falk and Greenbaum (1995); Haller and Krauss (2000), and Perezgonzalez (2014a), and discussed in, for example, Carver (1978); Nickerson (2000), Hubbard and Bayarri (2003); Kline (2004), and Goodman (2008). That they are still relevant in recent times testifies to the fact that the lessons of the past have not been learnt.

As the title anticipates, there is a twist to this saga, a pedagogical one: *p*-values are typically taught and presented as probabilities, and this may be the cause behind the confusions. A change in the heuristic we use for teaching and interpreting the meaning of *p*-values may be all we need to start working the path toward clarification and understanding.

In this article I will illustrate the differences in interpretation that a percentile heuristic and a probability one make. As guiding example, I will use a one-tailed *p*-value in a normal distribution— $z = -1.75$, $p = 0.04$; **Figure 1**). The default testing approach will be Fisher's tests of significance, but Neyman–Pearson's tests of acceptance approach will be assumed when discussing Type I errors and alternative hypotheses (for more information about those approaches see Perezgonzalez, 2014b, 2015). The scenario is the scoring of a sample of suspected schizophrenics on a validated psychological normality scale. The hypothesis tested (Fisher's H_0, Neyman–Pearson's H_M) is that the mean score of the sample on the normality scale does not differ from that of the normal population (no H_0 = the sample does not score as normal; H_A = the sample scores as schizophrenic, assuming previous knowledge that schizophrenics score low on the scale, by a given effect size). Neither a level of significance nor a rejection region is needed for the discussion.

FIGURE 1 | Location of an observed *z*-score and its corresponding *p*-value in the frequency distribution of the hypothesis under test. The accompanying scales are for the theoretical *z*-scores and percentiles, respectively.
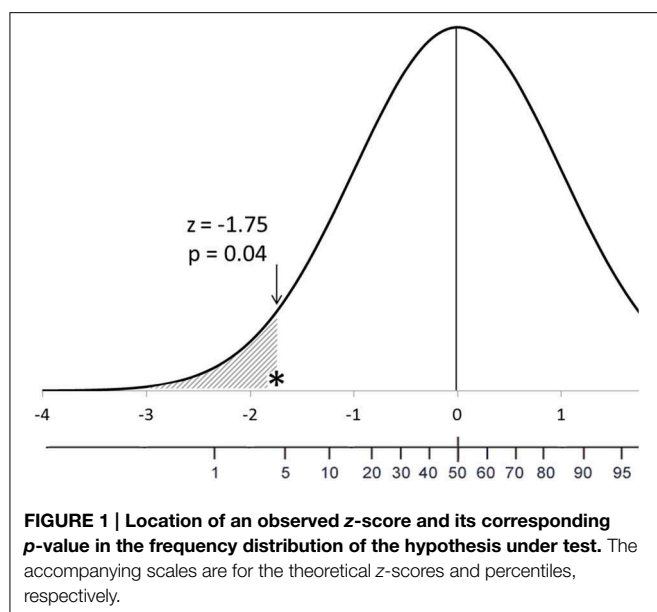
P-Values: Probabilities or Percentiles?

Let's start by establishing that *p*-values can be interpreted as probabilities. That is, when hypothetical population distributions are generated from sampling data, those frequency distributions follow the frequentist approach and the associated *p*-values show the appropriate probabilities. This is so because these *p*-values are theoretical—they represent the probability of, for example, a hypothetical human being alive today.

The *p*-value we obtain from our research data, however, is not a theoretical, probabilistic, value, but an observed one: its probability of occurrence is "1," precisely because it has occurred—it represents, for example, the realization that I am alive, not the probability of me being so. Therefore, the observed *p*-value does not represent a probability but a location in the distribution of reference. Among measures of location, percentiles (i.e., percentile ranks) are good heuristics to represent what observed *p*-values really are.

P-Values' Correct and Incorrect Misinterpretations

As **Figure 1** shows, a percentile describes a fact: the sample scored in the 4th percentile. As a probability, however, the *p*-value is often misinterpreted as, the observed result has a 4% likelihood of having occurred by chance—the *odds-against-chance fantasy* (Carver, 1978)—which also elicits a further misinterpretation as,

the observed result has a 96% likelihood of being a real effect (Kline, 2004).

The percentile heuristic also conveys the correct interpretation of the *p*-value as a cumulative percentage in the tail of the distribution: 4% of normal people will score this low or lower. As a probability, the *p*-value is often misinterpreted as, the sample has only a 4% likelihood of being normal—the *inverse probability error* (Cohen, 1994).

Consequently, because the percentile only provides information about location in the distribution of the normal scores hypothesis, it is impossible to know the probability of making a mistake if this hypothesis is rejected. As a probability, the *p*-value is often misinterpreted as, there is only a 4% likelihood of making a mistake when rejecting the tested hypothesis. This is further confused as, the probability of making a Type I error in the long run (alpha, α) is 4%; which then leads to the belief that α can be adjusted a posteriori—roving α (Goodman, 1993)—as a lower than anticipated Type I error (Kline, 2004; Perezgonzalez, 2015).

Furthermore, the percentile is circumscribed to its hypothesis of reference—normal scores on the normality test—and makes no concession for non-tested hypotheses. As a probability, the *p*-value is often misinterpreted as, there is a 96% likelihood that the sample scored as not normal—Fisher's negation of H_0, the *valid research hypothesis fantasy* (Carver, 1978)—or scored as schizophrenic—Neyman–Pearson's H_A, the *validity fallacy* (Mulaik et al., 1997).

Finally, the percentile heuristic helps ameliorate misinterpretations regarding future replicability, if only because we normally have enough experience with percentiles in other spheres of life as to realize that the big fish in this pond is neither necessarily big all the time nor equally big in all ponds. As a probability, the *p*-value is often misinterpreted as, there is a 96% likelihood that similar samples will score this low in future studies—the *replicability or reliability fallacy* (Carver, 1978).

Conclusions

The percentile heuristic is a more accurate model both for interpreting observed *p*-values and for preventing probabilistic misunderstandings. The percentile heuristic may also prove to be a better starting point for demystifying related statistical issues—such as the relationship among *p*-value, effect size and sample size—and epistemological issues—such as statistical significance, and the proving and disproving of hypotheses. All in all, the percentile heuristic matters for better statistical literacy and better research competence, allows for clearer understanding without imposing unnecessary cognitive workload, and has a positive effect in fostering the teaching and practice of psychological science.

References

Carver, R. P. (1978). The case against statistical significance testing. *Harv. Educ. Rev.* 48, 378–399.

Cohen, J. (1994). The earth is round (*p* <.05). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Falk, R., and Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theor. Psychol.* 5, 75–98. doi: 10.1177/0959354395051004

Goodman, S. N. (1993). *P* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137, 485–496.

Goodman, S. N. (2008). A dirty dozen: twelve P-value misconceptions. *Semin. Hematol.* 130, 995–1004. doi: 10.1053/j.seminhematol.2008.04.003

Haller, H., and Krauss, S. (2000). Misinterpretations of significance. A problem students share with their teachers. *MPR Online* 7, 1–20.

Hubbard, R., and Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Am. Stat.* 57, 171–178. doi: 10.1198/0003130031856

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: APA. doi: 10.1037/10693-000

Mulaik, S. A., Raju, N. S., and Harshman, R. (1997). "There is a time and place for significance testing," in *What If There were No Significance Tests?* eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah, NJ: Erlbaum), 65–116.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989X.5.2.241

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* Chichester: John Wiley & Sons.

Perezgonzalez, J. D. (2014a). Misinterpretation of the *p*-value and of the level of significance (2000). *Knowledge* 2014, 24–25. doi: 10.6084/m9.figshare.1320686

Perezgonzalez, J. D. (2014b). A reconceptualization of significance testing. *Theor. Psychol.* 24, 852–859. doi: 10.1177/0959354314546157

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* 6:223. doi: 10.3389/fpsyg.2015.00223

Schneider, J. W. (2015). Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 102, 411–432. doi: 10.1007/s11192-014-1251-5

Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing

Jose D. Perezgonzalez *

Business School, Massey University, Palmerston North, New Zealand

Despite frequent calls for the overhaul of null hypothesis significance testing (NHST), this controversial procedure remains ubiquitous in behavioral, social and biomedical teaching and research. Little change seems possible once the procedure becomes well ingrained in the minds and current practice of researchers; thus, the optimal opportunity for such change is at the time the procedure is taught, be this at undergraduate or at postgraduate levels. This paper presents a tutorial for the teaching of data testing procedures, often referred to as hypothesis testing theories. The first procedure introduced is Fisher's approach to data testing—tests of significance; the second is Neyman-Pearson's approach—tests of acceptance; the final procedure is the incongruent combination of the previous two theories into the current approach—NSHT. For those researchers sticking with the latter, two compromise solutions on how to improve NHST conclude the tutorial.

Keywords: test of significance, test of statistical hypotheses, null hypothesis significance testing, statistical education, teaching statistics, NHST, Fisher, Neyman-Pearson

INTRODUCTION

This paper introduces the classic approaches for testing research data: tests of significance, which Fisher helped develop and promote starting in 1925; tests of statistical hypotheses, developed by Neyman and Pearson (1928); and null hypothesis significance testing (NHST), first concocted by Lindquist (1940). This chronological arrangement is fortuitous insofar it introduces the simpler testing approach by Fisher first, then moves onto the more complex one by Neyman and Pearson, before tackling the incongruent hybrid approach represented by NHST (Gigerenzer, 2004; Hubbard, 2004). Other theories, such as Bayes's hypotheses testing (Lindley, 1965) and Wald's (1950) decision theory, are not object of this tutorial.

The main aim of the tutorial is to illustrate the bases of discord in the debate against NHST (Macdonald, 2002; Gigerenzer, 2004), which remains a problem not only yet unresolved but very much ubiquitous in current data testing (e.g., Franco et al., 2014) and teaching (e.g., Dancey and Reidy, 2014), especially in the biological sciences (Lovell, 2013; Ludbrook, 2013), social sciences (Frick, 1996), psychology (Nickerson, 2000; Gigerenzer, 2004) and education (Carver, 1978, 1993).

This tutorial is appropriate for the teaching of data testing at undergraduate and postgraduate levels, and is best introduced when students are knowledgeable on important background information regarding research methods (such as random sampling) and inferential statistics (such as frequency distributions of means).

In order to improve understanding, statistical constructs that may bring about confusion between theories are labeled differently, attending to their function in preference to their historical use (Perezgonzalez, 2014). Descriptive notes (notes) and caution notes (caution) are provided to clarify matters whenever appropriate.

FISHER'S APPROACH TO DATA TESTING

Ronald Aylmer Fisher was the main force behind tests of significance (Neyman, 1967) and can be considered the most influential figure in the current approach to testing research data (Hubbard, 2004). Although some steps in Fisher's approach may be worked out a priori (e.g., the setting of hypotheses and levels of significance), the approach is eminently inferential and all steps can be set up a posteriori, once the research data are ready to be analyzed (Fisher, 1955; Macdonald, 1997). Some of these steps can even be omitted in practice, as it is relatively easy for a reader to recreate them. Fisher's approach to data testing can be summarized in the five steps described below.

Step 1–Select an appropriate test. This step calls for selecting a test appropriate to, primarily, the research goal of interest (Fisher, 1932), although you may also need to consider other issues, such as the way your variables have been measured. For example, if your research goal is to assess differences in the number of people in two independent groups, you would choose a chi-square test (it requires variables measured at nominal levels); on the other hand, if your interest is to assess differences in the scores that the people in those two groups have reported on a questionnaire, you would choose a t-test (it requires variables measured at interval or ratio levels and a close-to-normal distribution of the groups' differences).

Step 2–Set up the null hypothesis (H_0). The null hypothesis derives naturally from the test selected in the form of an exact statistical hypothesis (e.g., H_0: M1–M2 = 0; Neyman and Pearson, 1933; Carver, 1978; Frick, 1996). Some parameters of this hypothesis, such as variance and degrees of freedom, are estimated from the sample, while other parameters, such as the distribution of frequencies under a particular distribution, are deduced theoretically. The statistical distribution so established thus represents the random variability that is theoretically expected for a statistical

nil hypothesis (i.e., $H_0 = 0$) given a particular research sample (Fisher, 1954, 1955; Bakan, 1966; Macdonald, 2002; Hubbard, 2004). It is called the null hypothesis because it stands to be nullified with research data (Gigerenzer, 2004).

Among things to consider when setting the null hypothesis is its directionality.

Directional and non-directional hypotheses. With some research projects, the direction of the results is expected (e.g., one group will perform better than the other). In these cases, a directional null hypothesis covering all remaining possible results can be set (e.g., H_0: M1–M2 = 0). With other projects, however, the direction of the results is not predictable or of no research interest. In these cases, a non-directional hypothesis is most suitable (e.g., H_0: M1–M2 = 0).

> **Notes:** H_0 does not need to be a nil hypothesis, that is, one that always equals zero (Fisher, 1955; Gigerenzer, 2004). For example, H_0 could be that the group difference is not larger than certain value (Newman et al., 2001). More often than not, however, H_0 tends to be zero.
>
> Setting up H_0 is one of the steps usually omitted if following the typical nil expectation (e.g., no correlation between variables, no differences in variance among groups, etc.). Even directional nil hypotheses are often omitted, instead specifying that one-tailed tests (see below) have been used in the analysis.

Step 3-Calculate the theoretical probability of the results under H_0 (p). Once the corresponding theoretical distribution is established, the probability (p-value) of any datum under the null hypothesis is also established, which is what statistics calculate (Fisher, 1955, 1960; Bakan, 1966; Johnstone, 1987; Cortina and Dunlap, 1997; Hagen, 1997). Data closer to the mean of the distribution (**Figure 1**) have a greater probability of occurrence under the null distribution; that is, they appear more frequently and show a larger p-value (e.g., $p = 0.46$, or 46 times in a 100 trials). On the other hand, data located further away from the mean have a lower probability of occurrence under the null distribution; that is, they appear less often and, thus, show a smaller p-value (e.g., $p = 0.003$). Of interest to us is the probability of our research results under such null distribution (e.g.,

the probability of the difference in means between two research groups).

The p-value comprises the probability of the observed results and also of any other more extreme results (e.g., the probability of the actual difference between groups and any other difference more extreme than that). Thus, the p-value is a cumulative probability rather than an exact point probability: It covers the probability area extending from the observed results toward the tail of the distribution (Fisher, 1960; Carver, 1978; Frick, 1996; Hubbard, 2004).

> **Note:** P-values provide information about the theoretical probability of the observed and more extreme results under a null hypothesis assumed to be true (Fisher, 1960; Bakan, 1966), or, said otherwise, the probability of the data given a true hypothesis—P(D|H); (Carver, 1978; Hubbard, 2004). As H_0 is always true (i.e., it shows the theoretical random distribution of frequencies under certain parameters), it cannot, at the same time, be false nor falsifiable a posteriori. Basically, if at any point you say that H_0 is false, then you are also invalidating the whole test and its results. Furthermore, because H_0 is always true, it cannot be proved, either.

Step 4–Assess the statistical significance of the results. Fisher proposed tests of significance as a tool for identifying research results of interest, defined as those with a low probability of occurring as mere random variation of a null hypothesis. A research result with a low p-value may, thus, be taken as evidence against the null (i.e., as evidence that it may not explain those results satisfactorily; Fisher, 1960; Bakan, 1966; Johnstone, 1987; Macdonald, 2002). How small a result ought to be in order to be considered statistically significant is largely dependent on the researcher in question, and may vary from research to research (Fisher, 1960; Gigerenzer, 2004). The decision can also be left to the reader, so reporting exact p-values is very informative (Fisher, 1973; Macdonald, 1997; Gigerenzer, 2004).

Overall, however, the assessment of research results is largely made bound to a given level of significance, by comparing whether the research p-value is smaller than such level of significance or not (Fisher, 1954, 1960; Johnstone, 1987):

- If the p-value is approximately equal to or smaller than the level of significance, the result is considered statistically significant.
- If the p-value is larger than the level of significance, the result is considered statistically non-significant.

Among things to consider when assessing the statistical significance of research results are the level of significance, and how it is affected by the directionality of the test and other corrections.

Level of significance (sig). The level of significance is a theoretical p-value used as a point of reference to help identify statistically significant results (**Figure 1**). There is no need to set up a level of significance a priori nor for a particular level of significance to be used in all occasions, although levels of significance such as 5% (sig ≈ 0.05) or 1% (sig ≈ 0.01) may be used for convenience, especially with novel research projects (Fisher, 1960; Carver, 1978; Gigerenzer, 2004). This highlights an important

FIGURE 1 | Location of a *t*-value and its corresponding *p*-value on a theoretical *t* distribution with 30 degrees of freedom. The actual p-value conveys stronger evidence against H_0 than sig ≈ 0.05 and can be considered highly significant.
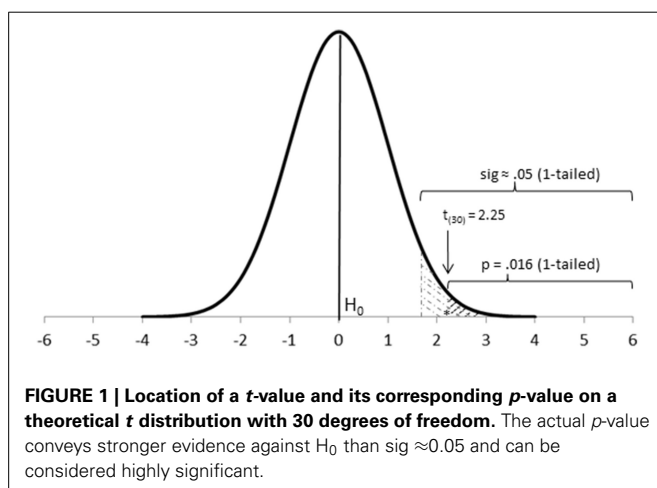
property of Fisher's levels of significance: They do not need to be rigid (e.g., *p*-values such as 0.049 and 0.051 have about the same statistical significance around a convenient level of significance of 5%; Johnstone, 1987).

Another property of tests of significance is that the observed *p*-value is taken as evidence against the null hypothesis, so that the smaller the *p*-value the stronger the evidence it provides (Fisher, 1960; Spielman, 1978). This means that it is plausible to gradate the strength of such evidence with smaller levels of significance. For example, if using 5% (sig ≈0.05) as a convenient level for identifying results which are just significant, then 1% (sig ≈0.01) may be used as a convenient level for identifying highly significant results and 1‰ (sig ≈0.001) for identifying extremely significant results.

> **Notes:** *Setting up a level of significance is another step usually omitted. In such cases, you may assume the researcher is using conventional levels of significance.*
>
> *If both H_0 and sig are made explicit, they could be joined in a single postulate, such as H_0: $M_1 - M_2 = 0$, sig ≈0.05.*
>
> *Notice that the p-value informs about the probability associated with a given test value (e.g., a t value). You could use this test value to decide about the significance of your results in a fashion similar to Neyman-Pearson's approach (see below). However, you get more information about the strength of the research evidence with p-values.*
>
> *Although the p-value is the most informative statistic of a test of significance, in psychology (e.g., American Psychological Association, 2010) you also report the research value of the test—e.g., $t_{(30)} = 2.25$, p = 0.016, 1-tailed. Albeit cumbersome and largely ignored by the reader, the research value of the test offers potentially useful information (e.g., about the valid sample size used with a test).*
>
> **Caution:** *Be careful not to interpret Fisher's p-values as Neyman-Pearson's Type I errors (α, see below). Probability values in single research projects are not the same than probability values in the long run (Johnstone, 1987), something illustrated by Berger (2003)—who reported that p = 0.05 often corresponds to α = 0.5 (or anywhere between α = 0.22 and α > 0.5)—and Cumming (2014)—who simulates the "dance" of p-values in the long run, commented further in Perezgonzalez (2015).*

One-tailed and two-tailed tests. With some tests (e.g., *F*-tests) research data can only be tested against one side of the null distribution (one-tailed tests), while other tests (e.g., *t*-tests) can test research data against both sides of the null distribution at the same time. With one-tailed tests you set the level of significance on the appropriate tail of the distribution. With two-tailed tests you cover both eventualities by dividing the level of significance between both tails (Fisher, 1960; Macdonald, 1997), which is commonly done by halving the total level of significance in two equal areas (thus covering, for example, the 2.5% most extreme positive differences and the 2.5% most extreme negative differences).

> **Note:** *The tail of a test depends on the test in question, not on whether the null hypothesis is directional or non-directional. However, you can use two-tailed tests as one-tailed ones when testing data against directional hypotheses.*

Correction of the level of significance for multiple tests. As we introduced earlier, a *p*-value can be interpreted in terms of its expected frequency of occurrence under the specific null distribution for a particular test (e.g., *p* = 0.02 describes a result that is expected to appear 2 times out of 100 under H_0). The same goes for theoretical *p*-values used as levels of significance. Thus, if more than one test is performed, this has the consequence of also increasing the probability of finding statistical significant results which are due to mere chance variation. In order to keep such probability at acceptable levels overall, the level of significance may be corrected downwards (Hagen, 1997). A popular correction is Bonferroni's, which reduces the level of significance proportionally to the number of tests carried out. For example, if your selected level of significance is 5% (sig ≈0.05) and you carry out two tests, then such level of significance is maintained overall by correcting the level of significance for each test down to 2.5% (sig ≈0.05/2 tests ≈0.025, or 2.5% per test).

> **Note:** *Bonferroni's correction is popular but controversial, mainly because it is too conservative, more so as the number of multiple tests increases. There are other methods for controlling the probability of false results when doing multiple comparisons, including familywise error rate methods (e.g., Holland and Copenhaver, 1987), false discovery rate methods (e.g., Benjamini and Hochberg, 1995), resampling methods (jackknifing, bootstrapping— e.g., Efron, 1981), and permutation tests (i.e., exact tests—e.g., Gill, 2007).*

Step 5–Interpret the statistical significance of the results. A significant result is literally interpreted as a dual statement: Either a rare result that occurs only with probability *p* (or lower) just happened, or the null hypothesis does not explain the research results satisfactorily (Fisher, 1955; Carver, 1978; Johnstone, 1987; Macdonald, 1997). Such literal interpretation is rarely encountered, however, and most common interpretations are in the line of "The null hypothesis did not seem to explain the research results well, thus we inferred that other processes—which we believe to be our experimental manipulation—exist that account for the results," or "The research results were statistically significant, thus we inferred that the treatment used accounted for such difference."

Non-significant results may be ignored (Fisher, 1960; Nunnally, 1960), although they can still provide useful information, such as whether results were in the expected direction and about their magnitude (Fisher, 1955). In fact, although always denying that the null hypothesis could ever be supported or established, Fisher conceded that non-significant results might be used for confirming or strengthening it (Fisher, 1955; Johnstone, 1987).

> **Note:** *Statistically speaking, Fisher's approach only ascertains the probability of the research data under a null hypothesis. Doubting or denying such hypothesis given a low p-value does not necessarily "support" or "prove" that the opposite is true (e.g., that there is a difference or a correlation in the population). More importantly, it does not "support" or "prove" that whatever else has been done in the research (e.g., the treatment*

used) explains the results, either (Macdonald, 1997). For Fisher, a good control of the research design (Fisher, 1955; Johnstone, 1987; Cortina and Dunlap, 1997), especially random allocation, is paramount to make sensible inferences based on the results of tests of significance (Fisher, 1954; Neyman, 1967). He was also adamant that, given a significant result, further research was needed to establish that there has indeed been an effect due to the treatment used (Fisher, 1954; Johnstone, 1987; Macdonald, 2002). Finally, he considered significant results as mere data points and encouraged the use of meta-analysis for progressing further, combining significant and non-significant results from related research projects (Fisher, 1960; Neyman, 1967).

HIGHLIGHTS OF FISHER'S APPROACH

Flexibility. Because most of the work is done a posteriori, Fisher's approach is quite flexible, allowing for any number of tests to be carried out and, therefore, any number of null hypotheses to be tested (a correction of the level of significance may be appropriate, though—Macdonald, 1997).

Better suited for ad-hoc research projects. Given above flexibility, Fisher's approach is well suited for single, *ad-hoc*, research projects (Neyman, 1956; Johnstone, 1987), as well as for exploratory research (Frick, 1996; Macdonald, 1997; Gigerenzer, 2004).

Inferential. Fisher's procedure is largely inferential, from the sample to the population of reference, albeit of limited reach, mainly restricted to populations that share parameters similar to those estimated from the sample (Fisher, 1954, 1955; Macdonald, 2002; Hubbard, 2004).

No power analysis. Neyman (1967) and Kruskal and Savage (Kruskal, 1980) were surprised that Fisher did not explicitly attend to the power of a test. Fisher talked about sensitiveness, a similar concept, and how it could be increased by increasing sample size (Fisher, 1960). However, he never created a mathematical procedure for controlling sensitiveness in a predictable manner (Macdonald, 1997; Hubbard, 2004).

No alternative hypothesis. One of the main critiques to Fisher's approach is the lack of an explicit alternative hypothesis (Macdonald, 2002; Gigerenzer, 2004; Hubbard, 2004), because there is no point in rejecting a null hypothesis without an alternative explanation being available (Pearson, 1990). However, Fisher considered alternative hypotheses implicitly—these being the negation of the null hypotheses—so much so that for him the main task of the researcher—and a definition of a research project well done—was to systematically reject with enough evidence the corresponding null hypothesis (Fisher, 1960).

NEYMAN-PEARSON'S APPROACH TO DATA TESTING

Jerzy Neyman and Egon Sharpe Pearson tried to improve Fisher's procedure (Fisher, 1955; Pearson, 1955; Jones and Tukey, 2000; Macdonald, 2002) and ended up developing an alternative approach to data testing. Neyman-Pearson's approach is more mathematical than Fisher's and does much of its work a priori, at the planning stage of the research project (Fisher, 1955; Macdonald, 1997; Gigerenzer, 2004; Hubbard, 2004). It also introduces a number of constructs, some of which are similar to those of Fisher. Overall, Neyman-Pearson's approach to

data testing can be considered tests of acceptance (Fisher, 1955; Pearson, 1955; Spielman, 1978; Perezgonzalez, 2014), summarized in the following eight main steps.

A PRIORI STEPS

Step 1–Set up the expected effect size in the population. The main conceptual innovation of Neyman-Pearson's approach was the consideration of explicit alternative hypotheses when testing research data (Neyman and Pearson, 1928, 1933; Neyman, 1956; Macdonald, 2002; Gigerenzer, 2004; Hubbard, 2004). In their simplest postulate, the alternative hypothesis represents a second population that sits alongside the population of the main hypothesis on the same continuum of values. These two groups differ by some degree: the effect size (Cohen, 1988; Macdonald, 1997).

Although the effect size was a new concept introduced by Neyman and Pearson, in psychology it was popularized by Cohen (1988). For example, Cohen's conventions for capturing differences between groups—d (**Figure 2**)—were based on the degree of visibility of such differences in the population: the smaller the effect size, the more difficult to appreciate such differences; the larger the effect size, the easier to appreciate such differences. Thus, effect sizes also double as a measure of importance in the real world (Nunnally, 1960; Cohen, 1988; Frick, 1996).

When testing data about samples, however, statistics do not work with unknown population distributions but with distributions of samples, which have narrower standard errors. In these cases, the effect size can still be defined as above because the means of the populations remain unaffected, but the sampling distributions would appear separated rather than overlapping (**Figure 3**). Because we rarely know the parameters of populations, it is their equivalent effect size measures in the context of sampling distributions which are of interest.

As we shall see below, the alternative hypothesis is the one that provides information about the effect size to be expected. However, because this hypothesis is not tested, Neyman-Pearson's procedure largely ignores its distribution except for a small percentage of it, which is called "beta" (β; Gigerenzer, 2004). Therefore, it is easier to understand Neyman-Pearson's procedure if we peg the effect size to beta and call it the expected minimum effect size (MES; **Figure 3**). This helps us conceptualize better how Neyman-Pearson's procedure works (Schmidt, 1996): The minimum effect size effectively represents that part of the main hypothesis that is not going to be rejected by the test (i.e., MES
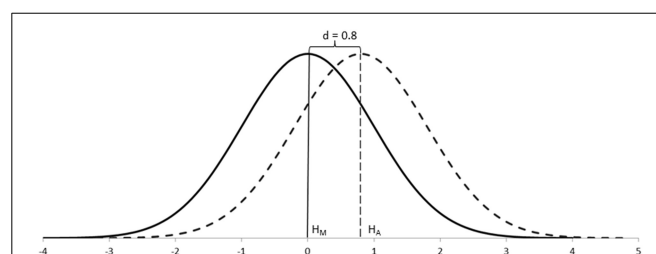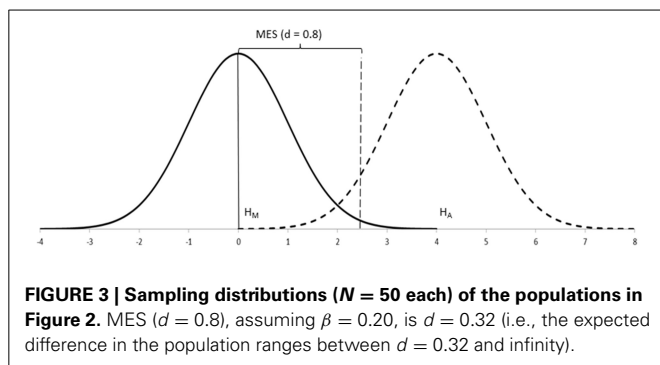

FIGURE 2 | A conventional large difference—Cohen's *d* = 0.8—between two normally distributed populations, as a fraction of one standard deviation.

FIGURE 3 | Sampling distributions ($N = 50$ each) of the populations in Figure 2. MES ($d = 0.8$), assuming $\beta = 0.20$, is $d = 0.32$ (i.e., the expected difference in the population ranges between $d = 0.32$ and infinity).

FIGURE 4 | Neyman-Pearson's approach tests data under H_M using the rejection region delimited by α. H_A contributes MES and β. Differences of research interest will be equal or larger than MES and will fall within this rejection region.

captures values of no research interest which you want to leave under H_M; Cortina and Dunlap, 1997; Hagen, 1997; Macdonald, 2002). (Worry not, as there is no need to perform any further calculations: The population effect size is the one to use, for example, for estimating research power.)

Note: A particularity of Neyman-Pearson's approach is that the two hypotheses are assumed to represent defined populations, the research sample being an instance of either of them (i.e., they are populations of samples generated by repetition of a common random process—Neyman and Pearson, 1928; Pearson, 1955; Hagen, 1997; Hubbard, 2004). This is unlike Fisher's population, which can be considered more theoretical, generated ad-hoc so as for providing the appropriate random distribution for the research sample at hand (i.e., a population of samples similar to the research sample—Fisher, 1955; Johnstone, 1987).

Step 2–Select an optimal test. As we shall see below, another of Neyman-Pearson's contributions was the construct of the power of a test. A spin-off of this contribution is that it has been possible to establish which tests are most powerful (for example, parametric tests are more powerful than non-parametric tests, and one-tailed tests are more powerful than two-tailed tests), and under which conditions (for example, increasing sample size increases power). For Neyman and Pearson, thus, you are better off choosing the most powerful test for your research project (Neyman, 1942, 1956).

Step 3–Set up the main hypothesis (H_M). Neyman-Pearson's approach considers, at least, two competing hypotheses, although it only tests data under one of them. The hypothesis which is the most important for the research (i.e., the one you do not want to reject too often) is the one tested (Neyman and Pearson, 1928; Neyman, 1942; Spielman, 1973). This hypothesis is better off written so as for incorporating the minimum expected effect size within its postulate (e.g., H_M: $M_1-M_2 = 0 \pm$ MES), so that it is clear that values within such minimum threshold are considered reasonably probable under the main hypothesis, while values outside such minimum threshold are considered as more probable under the alternative hypothesis (**Figure 4**).

Caution: Neyman-Pearson's H_M is very similar to Fisher's H_0. Indeed, Neyman and Pearson also called it the null hypothesis and often postulated it in a similar manner (e.g., as
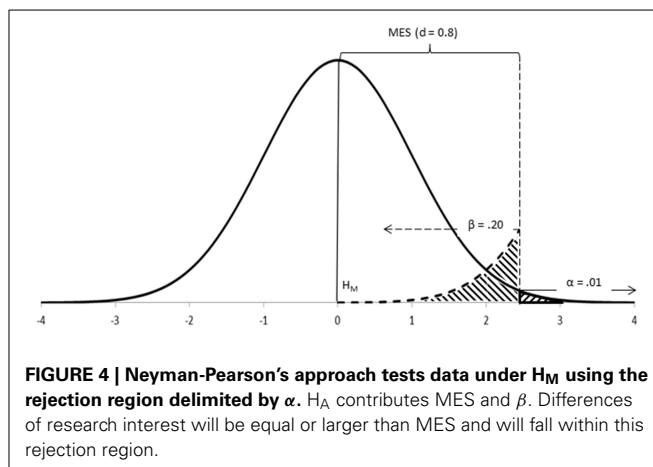
H_M: $M_1-M_2 = 0$). *However, this similarity is merely superficial on three accounts: H_M needs to be considered at the design stage (H_0 is rarely made explicit); it is implicitly designed to incorporate any value below the MES—i.e., the a priori power analysis of a test aims to capture such minimum difference (effect sizes are not part of Fisher's approach); and it is but one of two competing explanations for the research results (H_0 is the only hypothesis, to be nullified with evidence).*

The main aspect to consider when setting the main hypothesis is the Type I error you want to control for during the research.

Type I error. A Type I error (or error of the first class) is made every time the main hypothesis is wrongly rejected (thus, every time the alternative hypothesis is wrongly accepted). Because the hypothesis under test is your main hypothesis, this is an error that you want to minimize as much as possible in your lifetime research (Neyman and Pearson, 1928, 1933; Neyman, 1942; Macdonald, 1997).

Caution: A Type I error is possible under Fisher's approach, as it is similar to the error made when rejecting H_0 (Carver, 1978). However, this similarity is merely superficial on two accounts: Neyman and Pearson considered it an error whose relevance only manifests itself in the long run because it is not possible to know whether such an error has been made in any particular trial (Fisher's approach is eminently ad-hoc, so the risk of a long-run Type I error is of little relevance); therefore, it is an error that needs to be considered and minimized at the design stage of the research project in order to ensure good power—you cannot minimize this error a posteriori (with Fisher's approach, the potential impact of errors on individual projects is better controlled by correcting the level of significance as appropriate, for example, with a Bonferroni correction).

Alpha (α). Alpha is the probability of committing a Type I error in the long run (Gigerenzer, 2004). Neyman and Pearson often worked with convenient alpha levels such as 5% ($\alpha = 0.05$) and 1% ($\alpha = 0.01$), although different levels can also be set. The main hypothesis can, thus, be written so as for incorporating the alpha level in its postulate (e.g., H_M: $M_1-M_2 = 0\pm$ MES, $\alpha = 0.05$), to be read as the probability level at which

the main hypothesis will be rejected in favor of the alternative hypothesis.

> *Caution: Neyman-Pearson's α looks very similar to Fisher's sig. Indeed, Neyman and Pearson also called it the significance level of the test and used the same conventional cut-off points (5, 1%). However, this similarity is merely superficial on three accounts: α needs to be set a priori (not necessarily so under Fisher's approach); Neyman-Pearson's approach is not a test of significance (they are not interested in the strength of the evidence against H_M) but a test of acceptance (deciding whether to accept H_A instead of H_M); and α does not admit gradation—i.e., you may choose, for example, either α = 0.05 or α = 0.01, but not both, for the same test (while with Fisher's approach you can have different levels of more extreme significance).*

The critical region (CR_{test}) and critical value (CV_{test}, $Test_{crit}$) of a test. The alpha level helps draw a critical region, or rejection region (**Figure 4**), on the probability distribution of the main hypothesis (Neyman and Pearson, 1928). Any research value that falls outside this critical region will be taken as reasonably probable under the main hypothesis, and any research result that falls within the critical region will be taken as most probable under the alternative hypothesis. The alpha level, thus, also helps identify the location of the critical value of such test, the boundary for deciding between hypotheses. Thus, once the critical value is known—see below—, the main hypothesis can also be written so as for incorporating such critical value, if so desired (e.g., H_M: $M_1 - M_2 = 0 \pm$ MES, α = 0.05, $CV_t = 2.38$).

> *Caution: Neyman-Pearson's critical region is very similar to the equivalent critical region you would obtain by using Fisher's sig as a cut-off point on a null distribution. However, this similarity is rather unimportant on three accounts: it is based on a critical value which delimits the region to reject H_M in favor of H_A, irrespective of the actual observed value of the test (Fisher, on the contrary, is more interested in the actual p-value of the research result); it is fixed a priori and, thus, rigid and immobile (Fisher's level of significance can be flexible—Macdonald, 2002); and it is non-gradable (with Fisher's approach, you may delimit several more extreme critical regions as areas of stronger evidence).*

Step 4–Set up the alternative hypothesis (H_A). One of the main innovations of Neyman-Pearson's approach was the consideration of alternative hypotheses (Neyman and Pearson, 1928, 1933; Neyman, 1956). Unfortunately, the alternative hypothesis is often postulated in an unspecified manner (e.g., as H_A: $M_1 - M_2 \neq 0$), even by Neyman and Pearson themselves (Macdonald, 1997; Jones and Tukey, 2000). In practice, a fully specified alternative hypothesis (e.g., its mean and variance) is not necessary because this hypothesis only provides partial information to the testing of the main hypothesis (a.k.a., the effect size and β). Therefore, the alternative hypothesis is better written so as for incorporating the minimum effect size within its postulate (e.g., H_A: $M_1 - M_2 \neq 0 \pm$ MES). This way it is clear that values beyond such minimum effect size are the ones considered of research importance.

> *Caution: Neyman-Pearson's H_A is often postulated as the negation of a nil hypothesis (H_A: $M_1 - M_2 \neq 0$), which is coherent*

with a simple postulate of H_M (H_M: $M_1 - M_2 = 0$). These simplified postulates are not accurate and are easily confused with Fisher's approach to data testing—H_M resembles Fisher's H_0, and H_A resembles a mere negation of H_0. However, merely negating H_0 does not make its negation a valid alternative hypothesis—otherwise Fisher would have put forward such alternative hypothesis, something which he was vehemently against (Hubbard, 2004). As discussed earlier, Neyman-Pearson's approach introduces the construct of effect size into their testing approach; thus, incorporating such construct in the specification of both H_M and H_A makes them more accurate, and less confusing, than their simplified versions.

Among things to consider when setting the alternative hypothesis are the expected effect size in the population (see above) and the Type II error you are prepared to commit.

Type II error. A Type II error (or error of the second class) is made every time the main hypothesis is wrongly retained (thus, every time H_A is wrongly rejected). Making a Type II error is less critical than making a Type I error, yet you still want to minimize the probability of making this error once you have decided which alpha level to use (Neyman and Pearson, 1933; Neyman, 1942; Macdonald, 2002).

Beta (β). Beta is the probability of committing a Type II error in the long run and is, therefore, a parameter of the alternative hypothesis (**Figure 4**, Neyman, 1956). You want to make beta as small as possible, although not smaller than alpha (if β needed to be smaller than α, then H_A should be your main hypothesis, instead!). Neyman and Pearson proposed 20% (β = 0.20) as an upper ceiling for beta, and the value of alpha (β = α) as its lower floor (Neyman, 1953). For symmetry with the main hypothesis, the alternative hypothesis can, thus, be written so as for incorporating the beta level in its postulate (e.g., H_A: $M_1 - M_2 \neq 0 \pm$ MES, β = 0.20).

Step 5–Calculate the sample size (N) required for good power (1–β). Neyman-Pearson's approach is eminently a priori in order to ensure that the research to be done has good power (Neyman, 1942, 1956; Pearson, 1955; Macdonald, 2002). Power is the probability of correctly rejecting the main hypothesis in favor of the alternative hypothesis (i.e., of correctly accepting H_A). It is the mathematical opposite of the Type II error (thus, 1–β; Macdonald, 1997; Hubbard, 2004). Power depends on the type of test selected (e.g., parametric tests and one-tailed tests increase power), as well as on the expected effect size (larger ES's increase power), alpha (larger α's increase power) and beta (smaller β's increase power). A priori power is ensured by calculating the correct sample size given those parameters (Spielman, 1973). Because power is the opposite of beta, the lower floor for good power is, thus, 80% (1–β = 0.80), and its upper ceiling is 1–alpha (1–β = 1–α).

> *Note: H_A does not need to be tested under Neyman-Pearson's approach, only H_M (Neyman and Pearson, 1928, 1933; Neyman, 1942; Pearson, 1955; Spielman, 1973). Therefore, the procedure looks similar to Fisher's and, under similar circumstances (e.g., when using the same test and sample size), it will lead to the same results. The main difference between procedures is that Neyman-Pearson's H_A provides explicit information to the test; that is, information about ES and β. If this information is not*

taken into account for designing a research project with adequate power, then, by default, you are carrying out a test under Fisher's approach.

***Caution:** For Neyman and Pearson, there is little justification in carrying out research projects with low power. When a research project has low power, Type II errors are too big, so it is less probable to reject H_M in favor of H_A, while, at the same time, it makes unreasonable to accept H_M as the best explanation for the research results. If you face a research project with low a priori power, try the best compromise between its parameters (such as increasing α, relaxing β, settling for a larger ES, or using one-tailed tests; Neyman and Pearson, 1933). If all fails, consider Fisher's approach, instead.*

Step 6–Calculate the critical value of the test (CV_{test}, or Test$_{crit}$). Some of above parameters (test, α and N) can be used for calculating the critical value of the test; that is, the value to be used as the cut-off point for deciding between hypotheses (**Figure 5**, Neyman and Pearson, 1933).

A POSTERIORI STEPS

Step 7–Calculate the test value for the research (RV_{test}). In order to carry out the test, some unknown parameters of the populations are estimated from the sample (e.g., variance), while other parameters are deduced theoretically (e.g., the distribution of frequencies under a particular statistical distribution). The statistical distribution so established thus represents the random variability that is theoretically expected for a statistical main hypothesis given a particular research sample, and provides information about the values expected at different locations under such distribution.

By applying the corresponding formula, the research value of the test (RV_{test}) is obtained. This value is closer to zero the closer the research data is to the mean of the main hypothesis; it gets larger the further away the research data is from the mean of the main hypothesis.

***Note:** P-values can also be used for testing data when using Neyman-Pearson's approach, as testing data under H_M is similar to testing data under Fisher's H_0 (Fisher, 1955). It implies calculating the theoretical probability of the research data under the*

distribution of H_M—$P(D|H_M)$. Just be mindful that p-values go in the opposite way than RVs, with larger p-values being closer to H_M and smaller p-values being further away from it.

***Caution:** Because of above equivalence, you may use p-values instead of CV_{test} with Neyman-Pearson's approach. However, p-values need to be considered mere proxies under this approach and, thus, have no evidential properties whatsoever (Frick, 1996; Gigerenzer, 2004). For example, if working with a priori $\alpha = 0.05$, $p = 0.01$ would lead you to reject H_M at $\alpha = 0.05$; however, it would be incorrect to reject it at $\alpha = 0.01$ (i.e., α cannot be adjusted a posteriori), and it would be incorrect to conclude that you reject H_M strongly (i.e., α cannot be gradated). If confused, you are better off sticking to CV_{test}, and using p-values only with Fisher's approach.*

Step 8–Decide in favor of either the main or the alternative hypothesis. Neyman-Pearson's approach is rather mechanical once the a priori steps have been satisfied (Neyman and Pearson, 1933; Neyman, 1942, 1956; Spielman, 1978; Macdonald, 2002). Thus, the analysis is carried out as per the optimal test selected and the interpretation of results is informed by the mathematics of the test, following on the a priori pattern set up for deciding between hypotheses:

- If the observed result falls within the critical region, reject the main hypothesis and accept the alternative hypothesis.
- If the observed result falls outside the critical region and the test has good power, accept the main hypothesis.
- If the observed result falls outside the critical region and the test has low power, conclude nothing. (Ideally, you would not carry out research with low power—Neyman, 1955).

***Notes:** Neyman-Pearson's approach leads to a decision between hypotheses (Neyman and Pearson, 1933; Spielman, 1978). In principle, this decision should be between rejecting H_M or retaining H_M (assuming good power), as the test is carried out on H_M only (Neyman, 1942). In practice, it does not really make much difference whether you accept H_M or H_A, as appropriate (Macdonald, 1997). In fact, accepting either H_M or H_A is beneficial as it prevents confusion with Fisher's approach, which can only reject H_0 (Perezgonzalez, 2014).*

Reporting the observed research test value is relevant under Neyman-Pearson's approach, as it serves to compare the observed value against the a priori critical value—e.g., $t_{(64)} = 3.31$, 1-tailed > $CV_t = 2.38$, thus accept H_A. When using a p-value as a proxy for CV_{test}, simply strip any evidential value off p—e.g., $t_{(64)} = 3.31$, $p < \alpha$, 1-tailed.

Neyman-Pearson's hypotheses are also assumed to be true. H_M represents the probability distribution of the data given a true hypothesis—$P(D|H_M)$, while H_A represents the distribution of the data under an alternative true hypothesis—$P(D|H_A)$, even when it is never tested. This means that H_M and H_A cannot be, at the same time false, nor proved or falsified a posteriori. The only way forward is to act as if the conclusion reached by the test was true—subject to a probability α or β of making a Type I or Type II error, respectively (Neyman and Pearson, 1933; Cortina and Dunlap, 1997).
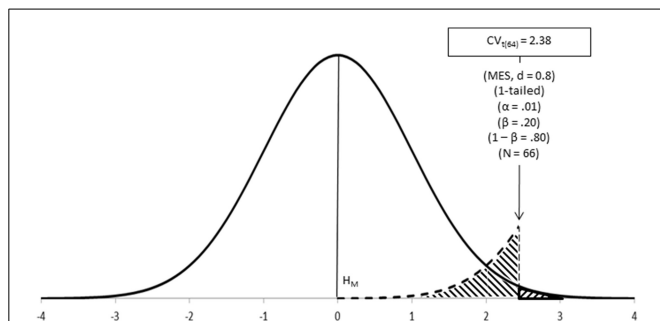
FIGURE 5 | Neyman-Pearson's test in action: CV_{test} is the point for deciding between hypotheses; it coincides with the cut-off points underlying α, β, and MES.

HIGHLIGHTS OF NEYMAN-PEARSON'S APPROACH

More powerful. Neyman-Pearson's approach is more powerful than Fisher's for testing data in the long run (Williams et al., 2006). However, repeated sampling is rare in research (Fisher, 1955).

Better suited for repeated sampling projects. Because of above, Neyman-Pearson's approach is well-suited for repeated sampling research using the same population and tests, such as industrial quality control or large scale diagnostic testing (Fisher, 1955; Spielman, 1973).

Deductive. The approach is deductive and rather mechanical once the a priori steps have been set up (Neyman and Pearson, 1933; Neyman, 1942; Fisher, 1955).

Less flexible than Fisher's approach. Because most of the work is done a priori, this approach is less flexible for accommodating tests not thought of beforehand and for doing exploratory research (Macdonald, 2002).

Defaults easily to Fisher's approach. As this approach looks superficially similar to Fisher's, it is easy to confuse both and forget what makes Neyman-Pearson's approach unique (Lehman, 1993). If the information provided by the alternative hypothesis—ES and β—is not taken into account for designing research with good power, data analysis defaults to Fisher's test of significance.

NULL HYPOTHESIS SIGNIFICANCE TESTING

NHST is the most common procedure used for testing data nowadays, albeit under the false assumption of testing substantive hypotheses (Carver, 1978; Nickerson, 2000; Hubbard, 2004; Hager, 2013). NHST is, in reality, an amalgamation of Fisher's and Neyman-Pearson's theories, offered as a seamless approach to testing (Macdonald, 2002; Gigerenzer, 2004). It is not a clearly defined amalgamation either and, depending on the author describing it or on the researcher using it, it may veer more toward Fisher's approach (e.g., American Psychological Association, 2010; Nunnally, 1960; Wilkinson and the Task Force on Statistical Inference, 1999; Krueger, 2001) or toward Neyman-Pearson's approach (e.g., Cohen, 1988; Rosnow and Rosenthal, 1989; Frick, 1996; Schmidt, 1996; Cortina and Dunlap, 1997; Wainer, 1999; Nickerson, 2000; Kline, 2004).

Unfortunately, if we compare Fisher's and Neyman-Pearson's approaches vis-à-vis, we find that they are incompatible in most accounts (**Table 1**). Overall, however, most amalgamations follow Neyman-Pearson procedurally but Fisher philosophically (Spielman, 1978; Johnstone, 1986; Cortina and Dunlap, 1997; Hubbard, 2004).

NHST is not only ubiquitous but very well ingrained in the minds and current practice of most researchers, journal editors and publishers (Spielman, 1978; Gigerenzer, 2004; Hubbard, 2004), especially in the biological sciences (Lovell, 2013; Ludbrook, 2013), social sciences (Frick, 1996), psychology (Nickerson, 2000; Gigerenzer, 2004) and education (Carver, 1978, 1993). Indeed, most statistics textbooks for those disciplines still teach NHST rather than the two approaches of Fisher and of Neyman and Pearson as separate and rather incompatible theories (e.g., Dancey and Reidy, 2014). NHST has also the (false) allure of being presented as a procedure for testing substantive hypotheses (Macdonald, 2002; Gigerenzer, 2004).

In the situations in which they are most often used by researchers, and assuming the corresponding parameters are also the same, both Fisher's and Neyman-Pearson's theories work with the same statistical tools and produce the same statistical results; therefore, by extension, NHST also works with the same statistical tools and produces the same results—in practice, however, both approaches start from different starting points and lead to different outcomes (Fisher, 1955; Spielman, 1978; Berger, 2003). In a nutshell, the differences between Fisher's and Neyman-Pearson's theories are mostly about research philosophy and about how to interpret results (Fisher, 1955).

The most coherent plan of action is, of course, to follow the theory which is most appropriate for purpose, be this Fisher's or Neyman-Pearson's. It is also possible to use both for achieving different goals within the same research project (e.g., Neyman-Pearson's for tests thought of a priori, and Fisher's for exploring the data further, a posteriori), pending that those goals are not mixed up.

However, the apparent parsimony of NHST and its power to withstand threats to its predominance are also understandable. Thus, I propose two practical solutions to improve NHST: the first a compromise to improve Fisher-leaning NHST, the second a compromise to improve Neyman-Pearson-leaning NHST. A computer program such as G*Power can be used for implementing the recommendations made for both.

IMPROVING FISHER-LEANING NHST

Fisher's is the closest approach to NHST; it is also the philosophy underlying common statistics packages, such as SPSS. Furthermore, because using Neyman-Pearson's concepts within NHST may be irrelevant or inelegant but hardly damaging, it requires little re-engineering. A clear improvement to NHST comes from incorporating Neyman-Pearson's constructs of effect size and of a priori sample estimation for adequate power. Estimating effect sizes (both a priori and a posteriori) ensures that researchers consider importance over mere statistical significance. A priori estimation of sample size for good power also ensures that the research has enough sensitiveness for capturing the expected effect size (Huberty, 1987; Macdonald, 2002).

IMPROVING NEYMAN-PEARSON-LEANING NHST

NHST is particularly damaging for Neyman-Pearson's approach, simply because the latter defaults to Fisher's if important constructs are not used correctly. An importantly damaging issue is the assimilation of p-values as evidence of Type I errors and the subsequent correction of alphas to match such p-values (roving α's, Goodman, 1993; Hubbard, 2004). The best compromise for improving NHST under these circumstances is to compensate a posteriori roving alphas with a posteriori roving betas (or, if so preferred, with a posteriori roving power). Basically, if you are adjusting alpha a posteriori (roving α) to reflect both the strength of evidence (sig) and the long-run Type I error (α), you should also adjust the long-run probability of making a Type II error (roving β). Report both roving alphas and roving betas for each test, and take them into account when interpreting your research results.

Table 1 | Equivalence of constructs in Fisher's and Neyman-Pearson's theories, and amalgamation of constructs under NHST.

Concept		Fisher		Neyman-Pearson
Test object		Data—$P(D\|H_0)$	$=$	Data—$P(D\|H_M)$
	NHST	➥	Data as if testing a falsifiable hypothesis—$P(H_0\|D)$	↩
Approach		A posteriori	$\neq$	A priori
	NHST	➥	A posteriori, sometimes both	↩ (partly)
Research goal		Statistical significance of research results	$\neq$	Deciding between competing hypotheses
	NHST	➥	Statistical significance, also used for deciding between hypotheses	↩
Hs under test		H_0, to be nullified with evidence	$\approx$	H_M, to be favored against H_A
	NHST	➥	Both ($H_0 = H_M$)	↩
Alternative hypothesis		Not needed (implicitly, "No H_0")	$\neq$	Needed. Provides ES and β
	NHST	➥	H_A posed as 'No H_0' (ES and β sometimes considered)	↩ (partly)
Prob. distr. of test		As appropriate for H_0	$=$	As appropriate for H_M
	NHST	➥	As appropriate for H_0	↩
Cut-off point		Sig identifies noteworthy results; can be gradated; can be corrected a posteriori	$\neq$	Common to CV_{test}, α, β, and MES; cannot be gradated; cannot be corrected a posteriori
	NHST	➥	Sig $= \alpha$, can be gradated, can be corrected a posteriori	↩ (partly)
Sample size calculator		None	$\neq$	Based on test, ES, α, and power $(1 - \beta)$
	NHST	➥	Either	↩
Statistic of interest		p-value, as evidence against H_0	$\neq$	CV_{test} (p-value has no inherent meaning but can be used as a proxy instead)
	NHST	➥	p-value, used both as evidence against H_0 and a proxy to accept H_A	↩
Error prob.		α possible, but irrelevant with single studies	$\neq$	$\alpha =$ Type I error prob. $\beta =$ Type II error prob.
	NHST	(partly) ➥	p-value $= \alpha =$ Type I error in single studies (β sometimes considered)	↩ (partly)
Result falls outside critical region		Ignore result as not significant	$\neq$	Accept H_M if good power; conclude nothing otherwise
	NHST	➥	Either ignore result as not significant; or accept H_0; or conclude nothing	↩
Result falls in critical region		Reject H_0	$\neq$	Accept H_A (= Reject H_M in favor of H_A)
	NHST	➥	Either	↩
Interpretation of results in critical region		Either a rare event occurred or H_0 does not explain the research data	$\neq$	H_A explains research data better than H_M does (given α)
	NHST		H_A has been proved / is true; or H_0 has been disproved / is false; or both	

(Continued)

Table 1 | Continued

Concept	Fisher		Neyman-Pearson
Next steps	Rejecting H_0 does not automatically justify not H_0. Replication needed, meta-analysis is useful.	$\neq$	Impossible to know whether α error has been made. Repeated sampling of same population needed, Monte Carlo is useful.
NHST		None (results taken as definitive, especially if significant); further studies may be sometimes recommended (especially if results are not significant)	

> **Caution:** NHST is very controversial, even if the controversy is not well known. A sample of helpful readings on this controversy are Christensen (2005); Hubbard (2004); Gigerenzer (2004); Goodman (1999), Louçã (2008, http://www.iseg.utl.pt/departamentos/economia/wp/wp022008deuece.pdf), Halpin and Stam (2006); Huberty (1993); Johnstone (1986), and Orlitzky (2012).

CONCLUSION

Data testing procedures represented a historical advancement for the so-called "softer" sciences, starting in biology but quickly spreading to psychology, the social sciences and education. These disciplines benefited from the principles of experimental design, the rejection of subjective probabilities and the application of statistics to small samples that Sir Ronald Fisher started popularizing in 1922 (Lehmann, 2011), under the umbrella of his tests of significance (e.g., Fisher, 1954). Two mathematical contemporaries, Jerzy Neyman and Egon Sharpe Pearson, attempted to improve Fisher's procedure and ended up developing a new theory, one for deciding between competing hypotheses (Neyman and Pearson, 1928), more suitable to quality control and large scale diagnostic testing (Spielman, 1973). Both theories had enough similarities to be easily confused (Perezgonzalez, 2014), especially by those less epistemologically inclined; a confusion fiercely opposed by the original authors (e.g., Fisher, 1955)—and ever since (e.g., Nickerson, 2000; Lehmann, 2011; Hager, 2013)—but something that irreversibly happened under the label of null hypothesis significance testing. NHST is an incompatible amalgamation of the theories of Fisher and of Neyman and Pearson (Gigerenzer, 2004). Curiously, it is an amalgamation that is technically reassuring despite it being, philosophically, pseudoscience. More interestingly, the numerous critiques raised against it for the past 80 years have not only failed to debunk NHST from the researcher's statistical toolbox, they have also failed to be widely known, to find their way into statistics manuals, to be edited out of journal submission requirements, and to be flagged up by peer-reviewers (e.g., Gigerenzer, 2004). NHST effectively negates the benefits that could be gained from Fisher's and from Neyman-Pearson's theories; it also slows scientific progress (Savage, 1957; Carver, 1978, 1993) and may be fostering pseudoscience. The best option would be to ditch NHST altogether and revert to the theories of Fisher and of Neyman-Pearson as—and when—appropriate. For everything else, there are alternative tools, among them exploratory data analysis (Tukey, 1977), effect sizes (Cohen, 1988), confidence intervals (Neyman, 1935), meta-analysis (Rosenthal, 1984), Bayesian applications (Dienes, 2014) and, chiefly, honest critical thinking (Fisher, 1960).

REFERENCES

American Psychological Association. (2010). *Publication Manual of the American Psychological Association*, 6th Edn. Washington, DC: APA.

Bakan, D. (1966). The test of significance in psychological research. *Psychol. Bull.* 66, 423–437. doi: 10.1037/h0020412

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57, 289–300.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Stat. Sci.* 18, 1–32. doi: 10.1214/ss/1056397485

Carver, R. P. (1978). The case against statistical significance testing. *Harv. Educ. Rev.* 48, 378–399. doi: 10.1080/00220973.1993.10806591

Carver, R. P. (1993). The case against statistical significance testing, revisited. *J. Exp. Educ.* 61, 287–292. doi: 10.1080/00220973.1993.10806591

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *Am. Stat.* 59, 121–126. doi: 10.1198/000313005X20871

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* New York, NY: Psychology Press.

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychol. Methods* 2, 161–172. doi: 10.1037/1082-989X.2.2.161

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Dancey, C. P., and Reidy, J. (2014). *Statistics Without Maths for Psychology, 6th Edn.* Essex: Pearson Education.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* 68, 589–599. doi: 10.1093/biomet/68.3.589

Fisher, R. A. (1932). Inverse probability and the use of likelihood. *Proc. Camb. Philos. Soc.* 28, 257–261. doi: 10.1017/S0305004100010094

Fisher, R. A. (1954). *Statistical Methods for Research Workers, 12th Edn.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 17, 69–78.

Fisher, R. A. (1960). *The Design of Experiments, 7th Edn.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference, 3rd Edn.* London: Collins Macmillan.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science* 345, 1502–1505. doi: 10.1126/science.1255484

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychol. Methods* 1, 379–390. doi: 10.1037/1082-989X.1.4.379

Gigerenzer, G. (2004). Mindless statistics. *J. Soc. Econ.* 33, 587–606. doi: 10.1016/j.socec.2004.09.033

Gill, P. M. W. (2007). Efficient calculation of p-values in linear-statistic permutation significance tests. *J. Stat. Comput. Simul.* 77, 55–61. doi: 10.1080/10629360500108053

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137, 485–496.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: the p-value fallacy. *Ann. Intern. Med.* 130, 995–1004. doi: 10.7326/0003-4819-130-12-199906150-00008

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *Am. Psychol.* 52, 15–24. doi: 10.1037/0003-066X.52.1.15

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: a methodological perspective. *Theor. Psychol.* 23, 251–270. doi: 10.1177/0959354312465483

Halpin, P. F., and Stam, H. J. (2006). Inductive inference or inductive behavior: fisher and Neyman–Pearson approaches to statistical testing in psychological research (1940-1960). *Am. J. Psychol.* 119, 625–653. doi: 10.2307/20445367

Holland, B. S., and Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43, 417–423. doi: 10.2307/2531823

Hubbard, R. (2004). Alphabet soup. Blurring the distinctions between p's and α's in psychological research. *Theor. Psychol.* 14, 295–327. doi: 10.1177/0959354304043638

Huberty, C. J. (1987). On statistical testing. *Educ. Res.* 8, 4–9. doi: 10.3102/0013189X016008004

Huberty, C. J. (1993). Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. *J. Exp. Educ.* 61, 317–333. doi: 10.1080/00220973.1993.10806593

Johnstone, D. J. (1986). Tests of significance in theory and practice. *Statistician* 35, 491–504. doi: 10.2307/2987965

Johnstone, D. J. (1987). Tests of significance following R. A. Fisher. *Br. J. Philos. Sci.* 38, 481–499. doi: 10.1093/bjps/38.4.481

Jones, L. V., and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychol. Methods* 5, 411–414. doi: 10.1037/1082-989X.5.4.411

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: APA.

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *Am. Psychol.* 56, 16–26. doi: 10.1037/0003-066X.56.1.16

Kruskal, W. (1980). The significance of Fisher: a review of R. A. Fisher: the life of a scientist. *J. Am. Stat. Assoc.* 75, 1019–1030. doi: 10.2307/2287199

Lehman, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypothesis: one theory or two? *J. Am. Stat. Assoc.* 88, 1242–1249. doi: 10.1080/01621459.1993.10476404

Lehmann, E. L. (2011). Fisher, Neyman, and the creation of classical statistics. New York, NY: Springer.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 1: Probability; Part 2: Inference.* Cambridge: Cambridge University Press.

Lindquist, E. F. (1940). *Statistical Analysis in Educational Research.* Boston, MA: Houghton Mifflin.

Louçã, F. (2008). *Should the Widest Cleft in Statistics - How and Why Fisher Opposed Neyman and Pearson.* WP/02/2008/DE/UECE, Technical University of Lisbon. Available online at: http://www.iseg.utl.pt/departamentos/economia/wp/wp022008deuece.pdf

Lovell, D. P. (2013). Biological importance and statistical significance. *J. Agric. Food Chem.* 61, 8340–8348. doi: 10.1021/jf401124y

Ludbrook, J. (2013). Should we use one-sided or two-sided p values in tests of significance? *Clin. Exp. Pharmacol. Physiol.* 40, 357–361. doi: 10.1111/1440-1681.12086

Macdonald, R. R. (1997). On statistical testing in psychology. *Br. J. Psychol.* 88, 333–347. doi: 10.1111/j.2044-8295.1997.tb02638.x

Macdonald, R. R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference. *Underst. Stat.* 1, 167–189. doi: 10.1207/S15328031US0103_03

Newman, I., Fraas, J., and Herbert, A. (2001). "Testing non-nil null hypotheses with T tests of group means: a monte carlo study," in *Annual Meeting Mid-Western Educational Research Association* (Chicago, IL).

Neyman, J. (1935). On the problem of confidence intervals. *Ann. Math. Stat.* 6, 111–116. doi: 10.1214/aoms/1177732585

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *J. R. Stat. Soc.* 105, 292–327. doi: 10.2307/2980436

Neyman, J. (1953). First *Course in Probability and Statistics.* New York, NY: Henry Holt.

Neyman, J. (1955). The problem of inductive inference. *Commun. Pure Appl. Math.* III, 13–46. doi: 10.1002/cpa.3160080103

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 18, 288–294.

Neyman, J. (1967). R. A. Fisher (1890-1962): an appreciation. *Science* 156, 1456–1460. doi: 10.1126/science.156.3781.1456

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 20A, 175–240. doi: 10.2307/2331945

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. A* 231, 289–337. doi: 10.1098/rsta.1933.0009

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989X.5.2.241

Nunnally, J. (1960). The place of statistics in psychology. *Educ. Psychol. Meas.* 20, 641–650. doi: 10.1177/001316446002000401

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organ. Res. Methods* 15, 199–228. doi: 10.1177/1094428111428356

Pearson, E. S. (1955). Statistical concepts in the relation to reality. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 17, 204–207.

Pearson, E. S. (1990). *'Student' a Statistical Biography of William Sealy Gosset.* Oxford: Claredon Press.

Perezgonzalez, J. D. (2014). A reconceptualization of significance testing. *Theor. Psychol.* 24, 852–859. doi: 10.1177/0959354314546157

Perezgonzalez, J. D. (2015). Confidence intervals and tests are two sides of the same research question. *Front. Psychol.* 6:34. doi: 10.3389/fpsyg.2015.00034

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research.* Beverly Hills, CA: Sage.

Rosnow, R. L., and Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *Am. Psychol.* 44, 1276–1284. doi: 10.1037/0003-066X.44.10.1276

Savage, I. R. (1957). Nonparametric statistics. *J. Am. Stat. Assoc.* 52, 331–344. doi: 10.1080/01621459.1957.10501392

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods* 1, 115–129. doi: 10.1037/1082-989X.1.2.115

Spielman, S. (1973). A refutation of the Neyman-Pearson theory of testing. *Br. J. Philos. Sci.* 24, 201–222. doi: 10.1093/bjps/24.3.201

Spielman, S. (1978). Statistical dogma and the logic of significance testing. *Philos. Sci.* 45, 120–135. doi: 10.1086/288784

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychol. Methods* 4, 212–213. doi: 10.1037/1082-989X.4.2.212

Wald, A. (1950). *Statistical Decision Functions.* Ney York, NY: Wiley.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. Guidelines and explanations. *Am. Psychol.* 54, 594–604. doi: 10.1037/0003-066X.54.8.594

Williams, R. H., Zimmerman, D. W., Ross, D. C., and Zumbo, B. D. (2006). *Twelve British Statisticians.* Raleigh, NC: Boson Books.

frontiers
in Psychology

CrossMark

Statistician, heal thyself: fighting statophobia at the source

Aleksandar Aksentijevic*

Department of Psychology, University of Roehampton, London, UK

Notwithstanding the popularity of psychology courses throughout the world, educators face a constant and difficult problem of overcoming the fear of and dislike for statistics which represents one of the pillars of modern psychological science. Although the issue is complex and multifaceted, here I argue that "statophobia" might represent a rational and justified response to the sense of unease felt in contact with abstract statistical concepts which are often vague, circular or ill-defined. I illustrate the problem by briefly discussing two myths about the nature of probability and statistics, namely that probability and statistics generate knowledge and that the fault for not understanding probability lies solely with the subjective cognition which is incapable of comprehending deeper mathematical truth. I argue that the confident presentation of statistical methods hides numerous conceptual blind spots that students might be aware of and that need to be addressed before other causes of statistics anxiety can be tackled successfully.

Keywords: statistics anxiety, randomness, variance, probability, information

WHO'S AFRAID OF THE BIG BAD… CENTRAL LIMIT THEOREM?

> Many candid persons, when confronted with the results of Probability, feel a strong sense of the uncertainty of the logical basis upon which it seems to rest. It is difficult to find an intelligible account of the meaning of "probability," or of how we are ever to determine the probability of any particular proposition; and yet treatises on the subject profess to arrive at complicated results of the greatest precision and the most profound practical importance (Keynes, 1921, p. 56).

Teaching statistics represents every psychology lecturer's baptism of fire. Facing a large auditorium packed with eager faces that start to sink into boredom and incomprehension as soon as the word "variance" is mentioned and its formula appears on the screen has filled many a new (as well as experienced) lecturer with a sense of foreboding and self-doubt. According to some estimates (e.g., Onwuegbuzie and Wilson, 2003), between 66 and 80% of students experience some degree of statistics anxiety.

Mathematics and statistics anxiety are related (Baloğlu, 1999) since statistics is formulated in the language of mathematics. Many of the causes of mathematics anxiety are transferrable to statistics, including difficulty of manipulating formulae as well as problems with performing arithmetical and algebraic operations. At the same time, research suggests that mathematics and statistics anxiety are distinct—if closely related—phenomena (Baloğlu, 2004). Some authors have observed utilization of different cognitive mechanisms (Cruise et al., 1985) and that statistical reasoning might be closer to verbal than mathematical reasoning (Buck, 1987). Like mathematics anxiety, statistics anxiety has been studied primarily using quantitative measures (e.g., STARS; Cruise et al., 1985). A number of dispositional and situational factors have been linked with statistics anxiety including gender, culture, tendency to procrastinate and reading ability (see Chew and Dillon, 2014a, for review).

Although some experts acknowledge the beneficial effects of medium anxiety levels (Keeley et al., 2008), statistics anxiety has been causally linked with reduced performance in a number

of disciplines—from psychology (Lalonde and Gardner, 1993; Macher et al., 2011) and education (Onwuegbuzie et al., 2000) to business (Zanakis and Valenzi, 1997). Consequently, a number of "treatments" has been proposed including reducing mathematical content and the amount of hand calculation, keeping students engaged, using humor, increasing instructor confidence and immediacy (Chew and Dillon, 2014a) and teaching online (DeVaney, 2010).

What could be causing statistics anxiety? Since a number of researchers cite negative attitudes toward statistics as the cause (e.g., Watson et al., 2002; Chiesi and Primi, 2010), the question should be rephrased—what causes the negativity (in addition to the factors mentioned above)? Statistics can be distinguished from mathematics in one important way—it aims to "freeze," quantify and package *uncertainty*—that fundamental imponderable of human existence. A recent systematic review of statistics anxiety literature (Chew and Dillon, 2014a) mentions only one study in which uncertainty features as a possible causal factor (Williams, 2013), and even there only as a psychological predisposition rather than an inherent property of statistics.

Although statistics teaching has come under increased scrutiny by the researchers, judging by the number of papers devoted to the topic in recent years, there is a creeping doubt that the problem lies not in the inability of students to "think properly" but in deep unresolved issues that underpin the foundations of probability and statistics. This is supported by the fact that expert researchers ostensibly exhibit an alarming lack of statistical aptitude—to the extent that the validity of most research findings in *most research fields* has been questioned (e.g., Ioannidis, 2005). Recurring episodes of heightened concern over statistical reasoning and performance of both students (current research topic) and experts (e.g., Cumming, 2014) suggest that the causes of anxiety and apprehension are at least partly to be found in the logic of statistical reasoning itself. Here, I briefly address two myths whose deconstruction might contribute to ameliorating the problem.

MYTH 1: STATISTICS GENERATES KNOWLEDGE

The development of powerful mathematical models and sophisticated inferential systems has engendered the belief that uncertainty is somehow controllable and—the worst of sins—conquerable. The ability to produce complex formulae which partition probabilities of various outcomes, weigh unequal conditional likelihoods and take into account prior knowledge does require mathematical sophistication that escapes many researchers, let alone students. At the same time, it fosters the mistaken impression that the formulae themselves generate qualitatively new information that is not present in the phenomena under observation.

If a pattern or a difference between objects is salient, our senses are sufficiently acute to detect and discriminate in most situations. Statistics becomes necessary when differences and dependencies become too small, numerous, complex, or remote to be analyzed by means of perception. This increase in informational distance between the observer and the phenomenon is managed via two basic steps. One is to exchange individual values/scores for a single number that hopefully retains maximum information. The second step is to quantify the uncertainty of this estimate. The amount of information conveyed by the mean is given by the variance. The more similar the scores, the lower the variance and the more informative the mean is. Here we face a paradox: The more informative the mean is, the less information there is in the population. To illustrate, in a population consisting of 4s, the mean of four conveys maximum information about the population. Yet, the population containing only 4s is maximally redundant and bereft of information (e.g., Shannon, 1948; Aksentijevic and Gibson, 2012). Thus, statistics provide most information about populations that possess no information at all. The more complex a data set, the less we can know about it. Rather than generating knowledge, statistics is at its best when no information is present.

The link between probabilistic models and real-life phenomena is tenuous at best. The use of probabilistic models in statistics is underpinned by a number of assumptions that can often not be confirmed empirically. Although this is dealt with by means of various methodological legerdemains, one example is sufficient to expose the students' predicament. In order for results of statistical tests to be interpreted in terms of a particular statistical model (e.g., Gaussian), we must assume that the process in question is unchanging over time (i.e., ergodic; see Attneave, 1959). Given the dynamic, ever-changing nature of reality on all scales, it is difficult to understand how the assumption of ergodicity can be maintained. If students cannot articulate these concerns, there is no reason to believe that they are not aware of them. Perhaps, anxiety stems from inchoate understanding of the impossibility of reconciling the fundamental unknowability of most future outcomes and the apparent certainty with which laws of probability and statistical procedures are expounded by experts.

The apparatus of statistical reasoning has its origins in the inability of scientists to describe and predict outcomes of complex processes—either on macro (gambling; Hacking, 1975) or micro scales (molecular motion; Uffink, 2006). Rather than a major advance in the search for truth, statistics could justifiably be viewed as an admission of defeat in the face of phenomena that defy easy description. Probabilistic reasoning can be reduced to the following statement: *In the absence of information about the process under observation*, all outcomes are equally likely—anything can happen. This statement is easily converted into a mathematical expression and elaborated in a number of ways to account for different combinations of outcomes. Equally, a posteriori probabilities can be modified by additional information (Bayesian calculus). However, none of these operations produces new information in the sense of affecting the reality on the ground. Probabilistic reasoning is a posteriori by definition. The best it can do is to roughly *describe* certain processes that are inaccessible to unaided perception.

MYTH 2: IT IS ALL OUR FAULT

An important contributory cause of statistics anxiety could be the constantly reinforced mantra, according to which human observers are failures at statistical reasoning (e.g., Kahneman and Tversky, 1972). This is in addition to apparent inability to reason logically (Wason, 1966) and well-documented biases observed in

simplest perceptual tasks such as bisecting a line (e.g., Jewell and McCourt, 2000). According to the dominant paradigm, human mind, that supposedly unique natural system replete with ability and potential is at the same time highly fallible and incapable of understanding even the basic tenets of logic and probability. If we combine this with the reluctance to question and challenge the teacher (Cruise et al., 1985), is it surprising that students feel anxious and uncomfortable from the start? Statistics anxiety could be more pernicious than mathematics anxiety. Mathematics is an enclosed system which exists independently of observation (although its subjective origins should be acknowledged). By contrast, probability makes inferences about real-life phenomena which all of us deal with regularly and understand intuitively. When told that our intuitions about our own experience are wrong, we are more likely to doubt our overall competence.

One of the most difficult problems encountered by lecturers is explaining the concept of randomness. As confirmed by the massive literature devoted to the subject, students are not the only ones that have difficulties with it. When "randomly" assigning subjects to conditions or generating "random" patterns, they might find that the supposedly random process often generates patterns that appear regular and repetitive (Lopes, 1982). Soon, they might experience a cognitive dissonance between the given definitions of randomness and their own intuitions. For instance, a random process involves infinity and complete independence between outcomes (e.g., Falk, 1991). How are we supposed to interact with a process that produces completely unknowable outcomes? When asked to generate "random-like" sequences, students soon learn that their performance systematically departs from the "laws" of probability. Specifically, they are told that they produce too many alternations and too few streaks (Gilovich et al., 1985; Oskarsson et al., 2009). And yet, if the "correct" distribution is known in advance, the process cannot be random. Students' anxiety might subside somewhat if they knew that mathematicians working for the RAND corporation were caught correcting tables of random numbers that were not sufficiently irregular (Gell-Mann, 1994).

The main source of confusion is the circular nature of "objective" probabilistic reasoning. Probabilistic models and ideas of randomness have subjective origins. Randomness represents abstract idealization of subjective complexity. Over time, it became so abstract as to lose any connection with its experiential sources. Randomization was invented in order to remove biases and preclude easy prediction. Randomization algorithms and other complex processes push the boundaries of complexity outside of the grasp of unaided perception and cognition. Is it then surprising that humans fail to understand randomness? Why would we expect humans whose cognition is pattern-based to be able to comprehend or generate sequences that lack any patterning or that conform to some probabilistic model? A random process can generate any outcome, leaving observers completely helpless. If they label a disordered sequence "random," they are told that this is no more random than a sequence of zeros, forcing them to suppress their (correct) intuition which says that ordered patterns are more likely to be generated by a deterministic process and that random patterns are generated by complex processes which they cannot understand. Equally, if they characterize an ordered

pattern as non-random, they are informed that they are wrong and that runs of identical symbols are often produced by random processes[1].

Related to this, one of the most consistent (and anxiety-inducing) findings in psychology has been the observation that subjects perform poorly on tasks requiring partitioning and weighting probabilities in the presence of partial information (Keren, 1984; Mandel, 2008). A good example is the three-card problem which produces significant departures from probabilistic norm (Falk and Lann, 2008). There are thee cards—red/red, red/green and green/green. If a card is drawn that shows a red face, what is the probability that its other face is red? A majority of subjects (at least 65%) failed to give a correct answer (2/3), preferring the uniform partitioning of probabilities (1/2)[2]. Following similar results obtained in related experiments, the authors concluded that "The size of the deviations from truth caused by falsely applying uniformity might not be practically pernicious, nonetheless, such judgments are *wrong in principle*. (p. 331; italics mine)" This sounds like an admonishment of the imperfect mind for its inability to keep up with the eternal mathematical truth. Yet, probabilistic calculus emerged from subjective observation and deduction. Following mathematical elaboration and abstraction, it became too detached from experience to remain relevant to reasoning about every-day events—for which purpose it had been invented in the first place. How can intuition, which created probability, be wrong when studied by its offspring? What matters is that having seen one red face, all we know (and can reasonably know) is that the second face could be either red or green. Knowing the correct probability tells us something about our long-term prospects of finding another red face *assuming that the uniformity decried by the authors is imposed on the sample space*, but nothing about what we are likely to find once we turn the card[3].

HONESTY IS THE BEST POLICY?

Statistics anxiety is a ubiquitous feature of social science courses. Part of the blame lies with the lack of practice, reputation of statistics as a "difficult" subject and mathematics-related issues. At the same time, learning to think statistically creates a conflict between intuition and the objective framework that constantly falsifies and challenges our understanding of how the world works. Although this is not necessarily wrong in itself, a closer inspection of probabilistic thinking shows that the counterintuitive nature of

---

[1]The fundamental disconnect between requirements of real-life research and randomness has caused a gradual weakening of the strict definition of the latter. Thus, Shannon (1948) speaks of a "known" random source and some authors have attempted to analyze the structure of random processes (e.g., Sun and Wang, 2010). Such attempts at "taming" randomness simply confirm the fundamental incompatibility between abstract probabilistic concepts and human perception and cognition (Aksentijevic, 2015).
[2]The sample consisted of over a 1000 students from an elite university.
[3]One of the greatest mathematicians of the twentieth century, Paul Erdös refused to accept the correct solution to the related "Monty Hall" problem (Vazsonyi, 1999). The solution depends on all prescribed possibilities being available equally often. This presumes uniformity—which is viewed as a fallacy when applied to individual outcomes. Also, see Keynes (1921, Chapter 5) on the impossibility of adjudging the truth of these alternative interpretations.

statistics does not originate in some deeper truth inaccessible to a lay observer, but is an unavoidable consequence of the dissonance between the fundamental limitations of human cognition and attempts to overcome these by means of mathematical formalism.

After years of training, some students conquer their anxiety and become proficient. As recommended in the literature, when facing the next generation of students, the newly fledged expert has to present a confident front and readily offer answers to difficult questions. But how can they maintain their confidence in the light of the finding that a large majority of expert researchers are well-nigh incompetent? In addition to focusing on putative antecedents of statistics anxiety, experts need to start a dialog that will shift the focus from the viability of various testing methods (e.g., the null-hypothesis significance testing, NHST) to discussing the appropriate role of statistics in research and more generally, science.

The first step could be to acknowledge the fundamental limitations of human mind and place statistics in this context. Rather than a panacea capable of advancing knowledge, probability and statistics should be viewed as an attempt to extend our informational reach into domains that are inherently beyond our grasp. We cannot know how successful our efforts are because the available tools are too simple to provide a complete (or even a partial) description of a phenomenon under observation. While being honest about limitations of statistics might not endear the lecturer to students who often crave certainty, honesty might pay off in the long run in terms of managing anxiety and unrealistic expectations as well as reducing the appeal of questionable practices. For if the relationship between statistics and reality is understood, more attention might be devoted to the psychological importance of experiments and less to the statistical significance of the result. At the same time, such a conceptual shift must be preceded by a substantial expert debate leading to a new consensus.

CONCLUSION

The ubiquitous problem of statistics anxiety has been investigated from many angles including gender (Rodarte-Luna and Sherry, 2008), motivation (Lavasani et al., 2014), and personality (Chew and Dillon, 2014b). However, none of the studies has considered that discomfort could partly originate in the disconnect between the certainty with which statistics is taught and the fundamental uncertainty inherent in it. This is of particular importance for psychologists who are expected to show a deeper understanding of the relationship between the mind and the statistical apparatus used to investigate it. In conclusion, I would like to offer the following summary which might reassure students next time they think they are incompetent because they do not understand probability and statistics:

a. Probability is an attempt to control uncertainty. It has no "laws"; it does not generate new information and has no predictive power. Ability to manipulate probabilities mathematically has no impact on individual outcomes of real-life processes. Predictability depends on the complexity of the phenomenon under observation and available resources. The more we know about the process, the more we know about its outcomes.

b. Statistics can help with extracting information from noise, but the trade-off is the increase in uncertainty with respect to interpretation. Rather than allowing us to "gain one up on the Universe," statistics is subject to the same fundamental cognitive limitations that necessitated its birth.

c. Probability models serve as *ad hoc* aids in framing research and providing reassurance and not as guarantors of truth. Outcomes of experiments might or might not come from a particular stochastic process but this can never be confirmed (or falsified). If someone objects that this does not affect the validity of statistical inference, the confused student would be justified in wondering why statistical models are used at all. The final arbiters of veridicality of a result are effect size (Cohen, 1969) and reproducibility[4]. These factors however say nothing about its importance.

d. Randomness is a mathematical idealization of subjective complexity. Students should be made aware that there is no such thing as a "random" process and that they are not in error when failing to reason or behave in accordance with probabilistic models. The problem lies elsewhere.

REFERENCES

Aksentijevic, A. (2015). No time for waiting: statistical structure reflects subjective complexity. *Proc. Natl Acad. Sci. U.S.A.* 112, E3159. doi: 10.1073/pnas.1507950112

Aksentijevic, A., and Gibson, K. (2012). Complexity and the cost of information processing. *Theor. Psychol.* 22, 572–590. doi: 10.1177/0959354311423435

Attneave, F. (1959). *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results*. New York, NY: Henry Holt.

Baloğlu, M. (1999). *A Comparison of Mathematics Anxiety and Statistics Anxiety in Relation to General Anxiety*. Available at: http://files.eric.ed.gov/fulltext/ED436703.pdf (accessed April 28, 2015)

Baloğlu, M. (2004). Statistics anxiety and mathematics anxiety: some interesting differences I. *Educ. Res. Q.* 27, 38–48.

Buck, J. L. (1987). More on superiority of women in statistics achievement: reply to Brooks. *Teach. Psychol.* 14, 45–47. doi: 10.1207/s15328023top1401_14

Chew, P. K. H., and Dillon, D. B. (2014a). Statistics anxiety update: refining the construct and recommendations for a new research agenda. *Perspect. Psychol. Sci.* 9, 196–208. doi: 10.1177/1745691613518077

Chew, P. K. H., and Dillon, D. B. (2014b). Statistics anxiety and the big five personality factors. *Procedia Soc. Behav. Sci.* 112, 1177–1186. doi: 10.1016/j.sbspro.2014.01.1282

Chiesi, F., and Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Stat. Educ. Res. J.* 9, 6–26.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cruise, R. J., Cash, R. W., and Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *Paper Presented at the Annual Meeting of the Statistical Education Section*, Chicago, IL.

Cumming, G. (2014). The new statistics: why and how? *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

---

[4]Even these are ultimately subjective. According to Cohen's (1969) convention, large effects are visible to the naked eye (visible difference). Reproducibility is another macroscopic subjective criterion (visible similarity).

DeVaney, T. A. (2010). Anxiety and attitude of graduate students in on-campus vs. online statistics courses. *J. Stat. Edu.* 18, 1–15.

Falk, R. (1991). Randomness—an ill defined but much needed concept. *J. Behav. Decis. Mak.* 4, 215–218.

Falk, R., and Lann, A. (2008). The allure of equality: uniformity in probabilistic and statistical judgment. *Cogn. Psychol.* 57, 293–334. doi: 10.1016/j.cogpsych.2008.02.002

Gell-Mann, M. (1994). *The Quark and the Jaguar: Adventures in the Simple and the Complex*. New York, NY: Freeman.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: on the misperception of random sequences. *Cogn. Psychol.* 17, 295–314. doi: 10.1016/0010-0285(85)90010-6

Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Jewell, G., and McCourt, M. E. (2000). Pseudoneglect: a review and meta-analysis of performance factors in line bisection tasks. *Neuropsychologia* 38, 93–100. doi: 10.1016/S0028-3932(99)00045-7

Kahneman, D., and Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cogn. Psychol.* 3, 430–454. doi: 10.1016/0010-0285(72)90016-3

Keeley, J., Zayac, R., and Correia, C. (2008). Curvilinear relationships between statistics anxiety and performance among undergraduate students: evidence for optimal anxiety. *Stat. Educ. Res. J.* 7, 4–15.

Keren, G. (1984). On the importance of identifying the correct 'problem space'. *Cognition* 16, 121–128.

Keynes, J. M. (1921). *A Treatise on Probability*. London: MacMillan and Company.

Lalonde, R. N., and Gardner, R. C. (1993). Statistics as a Second Language? A model for predicting performance in psychology students. *Can. J. Behav. Sci.* 25, 108–125. doi: 10.1037/h0078792

Lavasani, M. G., Weisani, M., and Shariati, F. (2014). The role of achievement goals, academic motivation in statistics anxiety. *Procedia Soc. Behav. Sci.* 114, 933–938. doi: 10.1016/j.sbspro.2013.12.810

Lopes, L. (1982). Doing the impossible: a note on induction and the experience of randomness. *J. Exp. Psychol. Learn.* 8, 626–636. doi: 10.1037/0278-7393.8.6.626

Macher, D., Paechter, M., Papousek, I., and Ruggeri, K. (2011). Statistics anxiety, trait anxiety, learning behavior, and academic performance. *Eur. J. Psychol. Educ.* 27, 483–498. doi: 10.1007/s10212-011-0090-5

Mandel, D. R. (2008). Violations of coherence in subjective probability: a representational and assessment process account. *Cognition* 106, 130–156. doi: 10.1016/j.cognition.2007.01.001

Onwuegbuzie, A. J., Slate, J. R., Paterson, F. R. A., Watson, M. H., and Schwartz, R. A. (2000). Factors associated with achievement in educational research courses. *Res. Schools* 7, 53–65.

Onwuegbuzie, A. J., and Wilson, V. A. (2003). Statistics anxiety: nature, etiology, antecedents, effects and treatments: a comprehensive review of the literature. *Teach. High. Educ.* 8, 195–209.

Oskarsson, A. T., Van Boven, McClelland, G., and Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychol. Bull.* 135, 262–285. doi: 10.1037/a0014821

Rodarte-Luna, B., and Sherry, A. (2008). Sex differences in the relation between statistics(anxiety)and cognitive/learning strategies. *Contemp. Educ. Psychol.* 33, 327–344. doi: 10.1016/j.cedpsych.2007.03.002

Shannon, C. (1948). A mathematical theory of communication. *Bell Tech. J.* 27, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Sun, Y., and Wang, H. (2010). Gambler's fallacy, hot hand belief, and the time of patterns. *Judgm. Decis. Mak.* 5, 124–132.

Uffink, J. (2006). "Compendium to the foundations of classical statistical physics," in *Handbook for the Philosophy of Physics*, eds J. Butterfield and J. Earman (Amsterdam: Elsevier), 924–1074.

Vazsonyi, A. (1999). Which door has the Cadillac? *Decis. Line* 30, 17–19.

Wason, P. C. (1966). "Reasoning," in *New Horizons in Psychology I*, ed. B. M. Foss (Harmondsworth: Penguin) 135–151.

Watson, F. S., Lang, T. R., and Kromrey, J. D. (2002). Breaking ground for EncStat: a statistics anxiety intervention program. *Paper Presented at the Annual Meeting of the Florida Educational Research Association*, Gainesville.

Williams, A. S. (2013). Worry, intolerance of uncertainty, and statistics anxiety. *Stat. Educ. Res. J.* 12, 48–59.

Zanakis, S. H., and Valenzi, E. R. (1997). Student anxiety and attitudes in business statistics. *J. Educ. Bus.* 73, 10–16. doi: 10.1080/08832329709601608

"Having to Shift Everything We've Learned to the Side": Expanding Research Methods Taught in Psychology to Incorporate Qualitative Methods

Lynne D. Roberts* and Emily Castell

School of Psychology and Speech Pathology, Curtin University, Perth, WA, Australia

In Australia the tradition of conducting quantitative psychological research within a positivist framework has been challenged, with calls made for the inclusion of the full range of qualitative and quantitative methodologies within the undergraduate psychology curriculum. Despite this, the undergraduate psychology curriculum in most Australian universities retains a strong focus on teaching quantitative research methods. Limited research has examined attitudes toward qualitative research held by undergraduate psychology students taught within a positivist framework, and whether these attitudes are malleable and can be changed through teaching qualitative methodologies. Previous research has suggested that students from strong quantitative backgrounds experience some cognitive dissonance and greater difficulties in learning qualitative methods. In this article we examine 3rd year undergraduate psychology students' attitudes to qualitative research prior to commencing and upon completion of a qualitative research unit. All students had previously completed two 13 weeks units of study in quantitative research methods. At Time 1, 63 students (84.1% female) completed online surveys comprising attitudinal measures. Key themes to emerge from student comments were that qualitative research was seen as an alternative approach, representing a paradigmatic shift that was construed by some students advantageous for meeting future professional and educative goals. Quantitative measures of attitudes to qualitative research were associated with general attitudes toward research, and psychology-specific epistemological beliefs. Changes in attitudes following completion of the qualitative research methods unit were in the hypothesized direction, but non-significant (small effect sizes). The findings increase our understanding of psychology students' attitudes toward qualitative research and inform our recommendations for teaching research methods within the undergraduate psychology curriculum.

Keywords: attitudes, undergraduate psychology, qualitative research, research methods, epistemological beliefs

INTRODUCTION

Qualitative research has a low profile in psychology, accounting for less than 10% of indexed empirical research articles published in psychology journals, with these publications predominantly in interdisciplinary and applied journals (Eagly and Riger, 2014). However, an increasing acceptance of the plurality of research methods (Gergen, 2014) has been accompanied by increased interest in qualitative research in psychology (O'Neill, 2002; Ponterotto, 2002, 2005; Karasz and Singelis, 2009; Demuth, 2015) as evidenced by the introduction of an American Psychological Association journal, *Qualitative Psychology*, specifically catering to presenting qualitative psychology findings (Gergen et al., 2015), and a small number of other journals explicitly encouraging the submission of qualitative papers (e.g., *The Journal of Counseling Psychology*; Haverkamp et al., 2005). Despite this, there remain wide differences in research practices and the knowledge and acceptability of qualitative research across sub-domains of psychology (Eagly and Riger, 2014), with quantitative psychological researchers often unaware of the range of qualitative epistemologies and practices available (Demuth, 2015).

In line with the increased interest in qualitative research amongst psychological researchers, qualitative research is increasingly being taught in undergraduate and postgraduate psychology degrees. For example, it is now mandatory for all undergraduate psychology courses in the United Kingdom to include qualitative research methods in the curriculum (Forrester and Koutsopoulou, 2008), although undergraduate research supervisors continue to report that the limited qualitative methods training provided in the undergraduate degree presents difficulties in supervising qualitative undergraduate dissertations (Wiggins et al., 2016). The tradition of conducting quantitative psychological research within a positivist framework is being challenged, with calls made for the inclusion of the full range of qualitative and quantitative methodologies within the undergraduate psychology curriculum (Mitchell et al., 2007; Breen and Darlaston-Jones, 2010; Wertz, 2014). Some universities, including our own, now teach qualitative and mixed methods research, in addition to quantitative methods, to undergraduate psychology students.

Despite the growth of qualitative methods in psychology and the teaching of qualitative methods in undergraduate and postgraduate psychology degrees, limited research has examined the attitudes toward qualitative research held by psychology students. Rabinowitz and Weseen (1997) explored how the quantitative-qualitative debate was experienced by 20 doctoral students in a social-personality psychology program. Most of the quantitatively oriented students expressed concerns that qualitative research was arbitrary, unscientific and particularly susceptible to researcher bias. These students also reported having difficulties evaluating qualitative studies. Murtonen (2005) examined social science, education and psychology students' preferences, aversions and appreciation of research methods and their readiness to use them. Students tended to have a dichotic attitude toward qualitative and quantitative research methods that was formed before or at the commencement of

their studies. Psychology students' interest in qualitative methods increased when they experienced difficulties in quantitative research. Mitchell et al. (2007) had three psychology students reflect on their experiences learning qualitative research methods as part of their undergraduate degree. One student acknowledged that they had internalized quantitative standards of research, such as external validity and objectivity, and described qualitative research as daunting as these standards appeared to be in direct opposition to the guiding principles of qualitative research. Students also reported that their exposure to qualitative methods was limited and that they had considerable difficulty obtaining the equipment necessary to conduct qualitative research.

More recently, Povee and Roberts (2014) interviewed 21 Australian psychology students and academics about their attitudes toward qualitative research. Qualitative research was seen by some participants as inherent to psychology, with parallels drawn between conducting qualitative research and practicing as a psychologist. Qualitative research methods were viewed as capturing the lived experience of research participants, reducing power differentials between the researcher and participants. However, qualitative research was viewed as less well respected and legitimate than quantitative methods within the field of psychology. Furthermore, viewing psychology in terms of a quantitative paradigm, participants raised concerns about the subjective nature of qualitative research, susceptibility to researcher bias, lack of rigor, inability to generalize beyond the sample and cast doubts about qualitative researchers' abilities in learning quantitative methods. Limited exposure to qualitative research methods and perceptions that qualitative research was time consuming and requiring large investments in resources were also identified as barriers to conducting qualitative research.

Many of the negative attitudes toward qualitative attitudes expressed in the literature reviewed may be a function of a lack of familiarity and training in qualitative methods. Despite the increasing prevalence of qualitative research in psychological research and education, limited research has examined how attitudes toward qualitative research change with teaching. Previous research has suggested that students from strong quantitative backgrounds experience some cognitive dissonance and greater difficulties in learning qualitative methods than other students (Kleinman et al., 1997; Cooper et al., 2012), resulting in challenges in learning "against the grain" (Eakin and Mykhalovskiy, 2005; Mitchell et al., 2007). Further, within the field of psychology, continuing resistance to qualitative research in some areas (McMullen, 2002) creates a context when methodological diversity may not be valued. Mitchell et al. (2007) highlight the importance of considering the epistemological beliefs of undergraduate psychology students upon commencing qualitative research education, noting the challenges involved in shifting epistemological beliefs.

Research in this area of attitudes has been hampered by the absence of reliable, validated measure of attitudes. However, the recent development and validation of measures of psychology specific epistemological beliefs (Renken et al., 2015) and attitudes toward qualitative research (Roberts and Povee, 2014; based on the qualitative research by Povee and Roberts, 2014), provide instruments suitable for this purpose.

The current study aims to explore undergraduate psychology students' attitudes to qualitative research, and how these change following exposure to qualitative methods. The context for the study is what has been described as a "typical academic study department in Australia" (Rees, 2013) where staff engage in research on a wide range of topics using a variety of research methodologies, including qualitative and mixed methods (Rees, 2013). Undergraduate psychology students complete two quantitative research methods units in their 2nd year and a qualitative methods unit and a mixed methods unit in their 3rd year of the degree.

The two research questions driving this research are:

(1) What attitudes do undergraduate psychology students hold toward qualitative research prior to commencing training in qualitative research?
(2) Do students' attitudes to qualitative research change after instruction in qualitative methods?

The first question is exploratory, designed to examine the relationship between general attitudes toward research, psychology specific epistemological beliefs and attitudes toward qualitative research. For the second research question, we hypothesized that instruction in qualitative methods would change attitudes to each of the four components of attitudes toward qualitative research:

H1. Perceptions of the lack of validity of qualitative research would decrease following instruction in qualitative methods.
H2. Perceptions that qualitative research captured the lived experience would increase following instruction in qualitative methods.
H3. Perceptions that qualitative search were time and resource intensive would increase following instruction in qualitative methods.
H4. Self-perceptions of a qualitative orientation would increase following instruction in qualitative methods.

MATERIALS AND METHODS

Participants

Sixty-three 3rd year undergraduate psychology students (84.1% female; age range 18–55 years) at an Australian university participated in this research. An *a priori* power analysis indicated a minimum sample size of 34 participants was required to have the power to detect a medium effect size change in attitudes.

Measures

Two online surveys were hosted on Qualtrics.com comprising the following measures:

Attitudes Toward Qualitative Research in Psychology (Roberts and Povee, 2014)

This measure consists of 18 items expressing attitudes toward qualitative research in psychology. Participants responded on a 7 point response scale ranging from strongly disagree (1)

to strongly agree (7). Exploratory and confirmatory factor analysis indicate four factors underlie the measure: 'perceived lack of validity' (example item, "Qualitative research lacks scientific rigor"), 'capturing the lived experience' (example item, "Qualitative research can capture the complexity of the social world"), 'qualitative orientation' (example item, "The most interesting findings in psychology are obtained with qualitative methods"), 'and 'time and resource intensive' (example item, "Qualitative research is harder to conduct than quantitative research"). The factors each have acceptable internal reliability as indicated by Cronbach's alpha: validity (0.82); capturing the lived experience (0.73), qualitative orientation (0.73) and time and resource intensive (0.72; Roberts and Povee, 2014).

Attitudes Toward Research Scale (Papanastasiou, 2005; Walker, 2010)

The original Attitudes Toward Research Scale (Papanastasiou, 2005) consisted of 32 items that measure general attitudes toward research. In this study we used the shortened version of the measure derived through confirmatory factor analysis (Walker, 2010) that comprises 18 items loading on three factors: research use (10 items), negative attributes of research (4 items), and positive attributes of research (4 items). Each item is responded to on a seven point scale anchored by strongly disagree (1) to strongly agree (7). Each factor has acceptable internal reliability (Cronbach's alphas all above 0.8; Walker, 2010).

Psychology-Specific Epistemological Beliefs Scale (Renken et al., 2015)

This scale comprises 13 items measuring psychology-specific epistemological beliefs. The items load onto three factors: significance of psychological research (example item "Carefully controlled research is not likely to be useful in solving psychological problems"), subjective nature of psychological knowledge (example item "Psychologists in different eras may use different theories and methods to interpret the same natural phenomenon"), and predictability of human behavior (example item "Psychological research can enable us to anticipate people's behavior with a high degree of accuracy"). Initial validation of this measure has included confirmatory factor analysis, test–retest reliability (correlations ranging between 0.65 and 0.78) and internal reliability ($\alpha = 0.54$ to 0.80 for subscales, and 0.75 to 0.82 for overall measure; Renken et al., 2015).

One open–ended question at Time 1 asked "How do you feel about completing a unit in Qualitative Research Methods? Why?"

Procedure

Following approval from Curtin University Human Research Ethics Committee, students enrolled in a compulsory 3rd year psychology undergraduate unit on Qualitative Methods at Curtin University were invited to take part in this research. Participation was voluntary, with students able to select from this and a range of other studies concurrently running through the School's research participation pool. The first survey was available for 2 weeks at the beginning of semester. Of the 190 students enrolled in the unit, 63 participated (33% response rate) in the Time 1 survey. In the last week of the semester participating students were emailed

a reminder to complete the second survey. Of these, 52 students completed the Time 2 survey. Students who completed both surveys were awarded participation points.

Data was downloaded from Qualtrics into SPSS (v. 20) for analysis. There were 13 missing data points in scale items in the Time 1 survey. Little's MCAR test indicated this data was missing completely at random ($\chi^2 = 75.017$, df = 891, $p = 1.000$), and the data points were replaced using Expectation Maximization. This dataset was used for exploring the first research question. There were also 13 missing data points in scale items in the Time 2 survey. Little's MCAR test indicated this data was missing completely at random ($\chi^2 = 0.000$, df = 860, $p = 1.000$), and the data points were replaced using Expectation Maximization. The Time 1 and 2 datasets were merged, with 35 cases able to be matched on the user-generated codes. This merged dataset was used for exploring the second research question.

RESULTS

Qualitative Results

To examine attitudes undergraduate psychology students hold toward qualitative research, response to the open–ended question were content analyzed.

Theme: An 'Alternative' Methodology

The theme 'an alternative approach' reflects students' tendency to frame their feelings about completing a unit in qualitative research methods in context of previously learned skills and information on quantitative research methods. A number of students suggested that the undergraduate curriculum had been dominated by quantitative research methods. It would seem that, for a number of students, feelings about undertaking a unit in qualitative research methods were informed by previous experiences in quantitative research methods units:

I am very excited to be learning about Qualitative Research methods. As I feel, that up and until now we have mainly focused on the Quantitative/Positivist aspect of Research through the use of statistics and experimental method.

Students expressed that they were looking forward to the prospect of learning approaches to research that were alternative to those used in quantitative research methods. A number of students articulated feeling apprehensive toward qualitative research methods, for example:

I feel intimidated because the content seems like such a contrast to the past research methods units I've completed, and I see it as a challenge.

In approaching the study of qualitative research methods, students appeared to construct qualitative methods as an 'alternative' approach to research, one that could be understood in terms of its differences and similarities to the dominant paradigm of quantitative research methods:

It should be interesting to compare what i already know about quantitative methods to qualitative and seeing not only the differences but their similarities.

The implication of this construction, perhaps, is reflected in discourse around the relative value of qualitative methods in contrast to quantitative methods. Perhaps the dominance of quantitative methods in the curriculum constructs an impression that the different research paradigms have a relative value attached to them:

"…we have had the importance of quantitative methods stressed to us so to oppose those methods and ways of thinking is overwhelming."

In considering how they felt about completing a unit in qualitative methods, some students noted that both qualitative and quantitative research methods were valuable, for example, one student noted:

"…both methods of research can be equally important in Psychology."

It would seem that attitudes toward learning about qualitative research methods were inextricably linked with previous learnings from units in quantitative research methods.

Theme: A Paradigmatic Shift

The theme 'a paradigmatic shift' captures students' reflections on epistemology in the context of learning about qualitative research methods. A number of students regarded qualitative research methods as demanding a different way of operating than required in previous quantitative research methods units, for example:

I am unsure about this unit. There is just some uncertainties I am yet to understand. Comparing this unit to my previous units, this is very theoretical…

Students' anticipated that learning about qualitative research methods would involve a level of uncertainty and ambiguity not previously encountered in their quantitative research methods units. For example:

I am a little apprehensive because it seems as if there are many gray areas within qualitative research and some aspects of qualitative research are not clearly defined.

A number of students aligned the shift from learning about quantitative methods to qualitative methods with a departure from focusing on numbers and statistics to exploring meaning and experiences. For example, when asked how they feel about the prospect of undertaking the unit, one student reflected:

"…I've never done anything like it before and nervous as I'm not very good with language/art topics but better with statistics."

Students indicated that the differences between quantitative and qualitative research methods reflected inherently different ways of approaching research and dealing with data, tantamount to 'wrapping' ones *"…head around a whole new set of ideas."* While a number of students expressed that they were apprehensive about undertaking a unit in qualitative methods, some students anticipated that qualitative research methods may offer a more intuitive way of approaching research than quantitative research methods:

"I just hope it makes a bit more real life sense than the stats we have completed so far!!!"

Some students noted the potential for qualitative research methods to offer depth, richness, and complexity, for example:

I believe that Qualitative Research Methods will allow me to incorporate my understanding of human's (both the subjects and researchers) complexities into the research, rather than attempting to remove our values, attitudes, and contexts from the experiments.

A number of students suggested that the emphasis on exploration, meaning, depth, and complexity inherent to qualitative research methods aligned with their personal interests, *"Excited to learn more about qualitative methods as that what I find interesting- opinions and beliefs people hold and the reasons behind them."* and with what they understand as the broader aims within the discipline of psychology:

It seems to tie in well with what I had in mind when I signed up to study Psych. I like that we have done quantitative methods first, it seems to ground this unit nicely in the realm of science.

Some students reflected on how they had been socialized to a particular approach to research methods, and anticipated that the alternative approach offered by qualitative research methods may pose a challenge to the dominant epistemological position fostered by previous units of study:

"It allows us to challenge our thinking and introduces different epistemological ideas/ theories. It will be exciting to see how our views get challenged over this year. . ."

Theme: Reconciling the 'Known' and 'Uncertain'

The theme captures a key tension emerging from students' feelings toward undertaking a unit in qualitative research methods. Students often reflected on their feelings toward undertaking the unit in terms of overarching goals. For example, some students expressed that undertaking the unit would be valuable for their future studies (e.g., *"I am very excited to begin qualitative research methods as I hope to be running this sort of research myself in future"*) and careers in psychology *"I feel like it is important to complete this unit as it will assist in my future career in psychology."*

For these students, learning about qualitative research methods seemed to be constructed as advantageous for meeting future professional and educative goals. While some students reflected on the unit as an opportunity to learn new information and enhance career goals, other students emphasized that completing the unit represented a necessary step in completing their degree. For example:

I'm not excited about completing the unit. It does not interest or stimulate me. However, I know it has to be done in order for me to get the most out of my degree and understand all elements and processes involved in psychological research.

Some students expressed indifference toward the content of the unit, expressing an eagerness to complete the unit and engage in future professional work:

I see the unit as a means to an end, the means to complete my degree in psychology and begin working in the field.

A number of students reflected on their feelings toward undertaking the unit in terms of how they felt this might impact upon their academic performance. Some students expressed apprehension toward the unit based on performance in previous research methods units. For example, those students who felt as though they had experienced difficulty in previous research methods units questioned their ability to perform well in qualitative research methods:

Initially I felt very distressed at the thought of completing this unit. Having struggled with previous Psychological Science units, I was anxious as I was unsure if I would find the unit more difficult and therefore, perhaps not pass it.

Other students expressed that the novelty of qualitative research method may pose a particular threat to academic success:

"I'm intrigued to find out what it's all about. I'm interested in learning a new way of approaching research questions, and a new way of thinking about knowledge and understanding, in general. I am nervous about the assignments and this unit, as its way outside my wheelhouse – I hope I don't bomb out and ruin my average and all the hard work I've put in so far."

For students who had experienced quantitative research methods as challenging, qualitative research methods offered an opportunity to learn a different approach which may offer an opportunity to perform:

The reason why I am interested in this unit is due to the fact that it is different to Quantitative research. In which using the computer for numbers was quiet confusing and slightly harder to grasp.

For some, the opportunity to embrace the uncertainty and novelty of qualitative research methods was appreciated, for others, the idea of undertaking a unit in qualitative research methods was seen as posing a challenge to academic performance, and potentially undermining their ability to do well in their studies.

Quantitative Results

Our first research questions asked what attitudes undergraduate psychology students hold toward qualitative research prior to commencing training in qualitative research. The scale scores and scale reliabilities from the Time 1 survey are presented in **Table 1**. On average, incoming students agreed that qualitative research captured the lived experience of participants and was time and resource intensive; however, they did not agree that qualitative research lacked validity or that they were qualitatively oriented.

To examine the first research question, scale measures of attitudes to qualitative research were correlated with measures of general attitudes toward research and psychology specific epistemological beliefs (see **Table 2**). Key findings in relation to general attitudes toward psychological research were that positive attitudes toward research were negatively associated with a qualitative orientation, and both positive attitudes and perceptions of research usefulness to the profession

TABLE 1 | Descriptive statistics for scale measures Time 1 ($N = 73$).

Measure	Range	Mean (SD)	Alpha
Attitudes toward qualitative research			
Perceived lack of validity	1.00–7.00	3.22 (1.38)	0.90
Capturing the lived experience	2.00–7.00	5.85 (1.09)	0.96
Time and resource intensive	1.50–6.25	4.33 (1.06)	0.76
Qualitative orientation	1.00–6.25	3.76 (1.05)	0.60
Attitudes toward Research			
Negative attributes of research	1.75–6.50	3.89 (1.17)	0.78
Positive attributes of research	1.75–7.00	4.62 (1.08)	0.74
Research use	3.40–7.00	5.70 (0.87)	0.89
Psychology specific epistemological beliefs			
Significance of psychological research	1.00–4.80	2.60 (0.90)	0.74
Subjective nature of psychological knowledge	3.60–6.80	5.38 (0.80)	0.79
Predictability of human behavior	3.33–7.00	4.87 (0.76)	0.66

TABLE 2 | Relationships between psychology specific epistemological beliefs, attitudes toward research and attitudes toward qualitative research ($N = 63$).

Measure	Attitudes to qualitative research			
	Lack of validity	Lived experience	Resource intensive	Qualitative orientation
Attitudes to research				
Negative	0.085	0.097	0.157	0.244
Positive	0.192	0.248*	0.212	−0.327**
Research use	0.169	0.355*	0.230	−0.136
Epistemological beliefs				
Significant	0.038	−0.194	−0.091	0.221
Subjective	−0.107	0.511**	0.078	0.027
Deterministic	−0.104	0.203	0.249*	−0.015

*$p < 0.05$, **$p < 0.01$, 2-tailed.

were positively correlated with viewing qualitative research as capturing the lived experience. Epistemological beliefs were also associated with attitudes toward qualitative research. In particular, perceptions of the subjective nature of psychological knowledge were positively associated with viewing qualitative research as capturing the lived experience.

To examine the second research question, four repeated measures *t*-tests were conducted on the four subscales of the *Attitudes toward Qualitative Research in Psychology* measure to test the four hypotheses. The mean scores and standard deviation for each scale at each time point are presented in **Table 3**. While all findings were in the hypothesized direction, there were no significant differences between Time 1 and Time 2 scores on the measures. Using Cohen's conventions, the effect sizes for each subscale except 'Time and Resource Intensive' were small.

DISCUSSION

The aim of our research was to explore undergraduate psychology students' attitudes to qualitative research, and how these change following exposure to training in qualitative research methods.

Prior to commencing training in qualitative research, we found students expressing mixed attitudes toward studying qualitative research methods. Based on the mean score for the 'qualitative orientation' subscale falling slightly below the mid-point of the sub-scale, on average students perceived themselves more strongly quantitatively than qualitatively oriented. The qualitative findings indicate that students viewed qualitative research methods as something 'other' than, and in opposition to, the quantitative research methods that they had been taught to date, with some students apprehensive about the prospect of learning a new methodological approach. These attitudes, represented in the theme 'an alternative methodology,' underpin a world view that qualitative and quantitative methods are dichotomous.

Viewing qualitative and quantitative methodologies as in opposition with each other is also reflected in the theme, 'a paradigmatic shift,' where student comments indicated that qualitative research was seen as more complex and interpretative than positivist quantitative methodologies, requiring a different way of thinking. The quantitative results indicate that epistemological beliefs about the subjective nature of psychological knowledge were strongly positively associated with the attitude that qualitative research captured the lived experience.

The third theme, 'reconciling the 'known' and 'uncertain,'' represents the goal-oriented views expressed by some students. Completing qualitative methods training was seen as beneficial to future studies, to completing their degree and to future work. Deterministic beliefs about the predictability of human behavior were associated with beliefs that qualitative research was (unnecessarily) time and resource intensive, presumably in comparison to quantitative research. The uncertainty about qualitative methods is also captured in the greater variance in scores on the 'perceived lack of validity' subscale of the Attitudes to Qualitative research measure, in comparison to other subscales of the same measure, indicating the greater divergence of views about the (in)validity of qualitative research.

The dominance of quantitative methods in the first 2 years of the undergraduate curriculum constructs a tension for students when approaching qualitative research methods. While students expressed that they were looking forward to learning what qualitative approaches to research could offer beyond those advantages offered by quantitative methods, for some, undertaking qualitative research methods posed a threat to their performance and prior learnings. The teaching of quantitative methods prior to qualitative methods sets quantitative methods

TABLE 3 | Pre and post-test scores on attitudes toward qualitative research scale ($N = 35$).

	Pre-test Mean (SD)	Post-test Mean (SD)	Effect size Cohen's *d*
Perceived lack of validity	3.29 (1.48)	3.00 (1.22)	0.22
Capturing the lived experience	5.88 (0.85)	6.09 (0.88)	0.25
Time and resource intensive	4.19 (1.09)	4.23 (0.98)	0.04
Qualitative orientation	3.78 (1.11)	4.10 (1.28)	0.27

up as the main-stream, preferred research orientation in psychology. This focus on quantitative methodologies, often with limited reference to the underlying epistemological values, positions qualitative research as the alternative, and 'lesser' methodological paradigm.

The privileging of quantitative methods in undergraduate psychology education may be in contrast to the expectations of students electing to study psychology. On average, undergraduate psychology students have greater interest in practitioner than research activities (Holmes and Beins, 2009; Holmes, 2014). In summarizing the literature on teaching introductory research methods Earley (2014) noted that across disciplines (including psychology) students have misconceptions about research, see little relevance of research methods to their planned future careers, and may lack interest and motivation. The high mean score on the 'capturing the lived experience scale' of the Attitudes Toward Qualitative Research measure in this research suggests that despite socialization into psychology as a quantitative science, many students continue to see value in qualitative approaches, even after quantitative research training.

When reoriented to qualitative methods in the 3rd year of their undergraduate psychology degree, students may experience some dissonance between their (post)positivist quantitative methods training with the emphasis on control, rigor and generalisability and the competing values (including the embracing of exploration, subjectivity, and experience) associated with qualitative methods and their conceptions of psychology upon entering the degree. Students identified that there is a level of uncertainty and ambiguity in qualitative research that they had not previously encountered in quantitative research units, that the qualitative approach may challenge them and require some shifts in thinking. Students may initially struggle with holding multiple ways of knowing simultaneously in mind, and how to integrate their thinking about these. This process occurs in the context of students ultimately striving to 'perform' and achieve good grades, potentially restricting 'deep' or meaningful learning where they embrace the unknown and the risk of 'getting it wrong.' These findings are consistent with previous findings indicating that students from strong quantitative backgrounds experience dissonance when faced with qualitative approaches (Kleinman et al., 1997; Cooper et al., 2012). Integrating qualitative research methods against a backdrop of what they have learned about quantitative methods, qualitative methods inevitably become the 'alternative' to quantitative methods.

We were interested in whether students' attitudes to qualitative research changed after instruction in qualitative methods. We found that completion of the qualitative research methods unit resulted in small increases in attitudes toward qualitative research in the hypothesized direction, but these shifts were not statistically significant. It is possible that attitudes toward research are already largely 'set' following socialization into psychology as a (quantitative) science and are resistant

to change. Difficulties in in learning "against the grain" have been reported previously (Eakin and Mykhalovskiy, 2005; Mitchell et al., 2007). Further, increasing knowledge of research methods does not necessarily result in increased positive attitudes (Sizemore and Lewandowski, 2009). However, the current research was limited by the small sample size, high attrition rates and difficulties in matching pre- and post- responses. Being asked to take part in study may also have contributed to the perception that qualitative methods are different, an inevitable aspect of research in this area. Further research using larger samples is required to more fully test the malleability of psychology students' attitudes toward qualitative research through education. Individual case studies may provide insights into how student perceptions change through experiencing qualitative research training. Those prospective challenges identified by students were based on anticipation, as opposed to engagement with, and reflection upon, experiences with qualitative research methods training. Exploring students' reflections on their engagement in qualitative research methods training after-the-fact may give further insight into the difficulties encountered in in practice.

The ordering of teaching quantitative and qualitative research may be an important consideration in shaping students' orientation toward the full range of research methods. The current focus on teaching quantitative research before qualitative research privileges quantitative research and sets qualitative research as the 'alternative' methodology. This curriculum structure is perhaps not conducive to fostering in students an appreciation of the methodological diversity (McMullen, 2002) which is valued in contemporary approaches to psychological research. Embedding teaching of the epistemological foundations of psychology (Breen and Darlaston-Jones, 2010) and the full range of methods and methodologies available from the start of the undergraduate psychology degree may help to legitimize qualitative findings and position qualitative research as valued within psychology (Gough and Lyons, 2016). It may also serve to remove the false dichotomy between qualitative and quantitative methods and lay the foundations for future mixed methods research.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENT

REFERENCES

Breen, L. J., and Darlaston-Jones, D. (2010). Moving beyond the enduring dominance of positivism in psychological research: implications for psychology in Australia. *Aust. Psychol.* 45, 67–76. doi: 10.1080/00050060903127481

Cooper, R., Chenail, R. J., and Fleming, S. (2012). A grounded theory of inductive qualitative research education: results of a meta-data-analysis. *Qual. Rep.* 17, 8, 1–26

Demuth, C. (2015). New directions in qualitative research in psychology. *Integr. Psychol. Behav. Sci.* 49, 125–133. doi: 10.1007/s12124-015-9303-9

Eagly, A. H., and Riger, S. (2014). Feminism and psychology: critiques of methods and epistemology. *Am. Psychol.* 59, 685–702. doi: 10.1037/a0037372

Eakin, J. M., and Mykhalovskiy, E. (2005). Teaching against the grain: the challenges of teaching qualitative research in the health sciences. *Forum. Qual. Soc. Res.* 6:42.

Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teach. High. Educ.* 19, 242–253. doi: 10.1080/13562517.2013.860105

Forrester, M. A. and Koutsopoulou, G. Z. (2008). Providing resources for enhancing the teaching of qualitative methods at the undergraduate level: current practices and the work of the HEA Psychology Network Group. *Qual. Res. Psychol.* 5, 173–178. doi: 10.1080/14780880802314312

Gergen, K. J. (2014). Pursuing excellence in qualitative inquiry. *Qual. Psychol.* 1, 49–60. doi: 10.1037/qup0000002

Gergen, K. J., Josselson, R., and Freeman, M. (2015). The promises of qualitative inquiry. *Am. Psychol.* 70, 1–9. doi: 10.1037/a0038597

Gough, B., and Lyons, A. (2016). The future of qualitative research in psychology: accentuating the positive. *Integr. Psychol. Behav. Sci.* 50, 234–243. doi: 10.1007/s12124-015-9320-8

Haverkamp, B. E., Morrow, S. L., and Ponterotto, J. G. (2005). A time and a place for qualitative and mixed methods in counseling psychology research. *J. Couns. Psychol.* 52, 123–125. doi: 10.1037/0022-0167.52.2.123

Holmes, J. D. (2014). Undergraduate psychology's scientific identity dilemma: student and instructor interests and attitudes. *Teach. Psychol.* 41, 104–109. doi: 10.1177/0098628314530339

Holmes, J. D., and Beins, B. C. (2009). Psychology is a science: at least some students think so. *Teach. Psychol.* 36, 5–11. doi: 10.1080/00986280802529350

Karasz, A., and Singelis, T. M. (2009). Qualitative and mixed methods research in cross-cultural psychology. *J. Cross-Cult. Psychol.* 40, 909–916. doi: 10.1177/0022022109349172

Kleinman, S., Vopp, M. A., and Henderson, K. A. (1997). Qualitatively different: teaching fieldwork to graduate students. *J. Contemp. Ethnogr.* 25, 469–499. doi: 10.1177/089124197025004003

McMullen, L. M. (2002). Learning the languages of research: transcending illiteracy and indifference. *Can. Psychol.* 43, 195–204. doi: 10.1037/h0086916

Mitchell, T., Friesen, M., Friesen, D., and Rose, R. (2007). Learning against the grain: reflections on the challenges and revelations of studying qualitative research methods in an undergraduate psychology course. *Qual. Res. Psychol.* 4, 227–240. doi: 10.1080/14780880701473441

Murtonen, M. (2005). University students' research orientations: do negative attitudes exist toward quantitative methods? *Scand. J. Educ. Res.* 49, 263–280. doi: 10.1080/00313830500109568

O'Neill, P. (2002). Tectonic change: the qualitative paradigm in psychology. *Can. Psychol.* 43, 190–194. doi: 10.1037/h0086915

Papanastasiou, E. C. (2005). Factor structure of the "Attitudes Toward Research" scale. *Stat. Educ. Res. J.* 4, 16–26

Ponterotto, J. G. (2002). Qualitative research methods: the fifth force in psychology. *Coun. Psychol.* 30, 394–406. doi: 10.1177/0011000002303002

Ponterotto, J. G. (2005). Qualitative research in counseling psychology: a primer on research paradigms and philosophy of science. *J. Coun. Psychol.* 52, 126–136. doi: 10.1037/0022-0167.52.2.126

Povee, K., and Roberts, L. D. (2014). Qualitative research in psychology: attitudes of psychology students and academic staff. *Aust. J. Psychol.* 66, 28–37. doi: 10.1111/ajpy.12031

Rabinowitz, V. C., and Weseen, S. (1997). Elu(ci)d(at)ing epistemological impasses: re-viewing the qualitative/quantitative debates in psychology. *J. Soc. Issues* 53, 605–630. doi: 10.1111/j.1540-4560.1997.tb02451.x

Rees, C. S. (2013). Promoting psychology to students: embracing the multiplicity of research foci and method. *Front. Psychol.* 4:774. doi: 10.3389/fpsyg.2013.00774

Renken, M. D., McMahan, D. A., and Nitkova, M. (2015). Initial validation of an instrument measuring psychology-specific epistemological beliefs. *Teach. Psychol.* 42, 126–136. doi: 10.1177/0098628315569927

Roberts, L. D., and Povee, K. (2014). A brief measure of attitudes towards qualitative research in psychology. *Aust. J. Psychol.* 66, 249–256. doi: 10.1111/ajpy.12059

Sizemore, O. J. and Lewandowski, G. W. (2009). Learning might not equal liking: research methods course changes knowledge but not attitudes. *Teach. Psychol.* 36, 90–95. doi: 10.1080/00986280902739727

Walker, D. A. (2010). A confirmatory factor analysis of the attitudes toward research scale. *Mult. Linear Regression Viewp.* 36, 18–27.

Wertz, F. J. (2014). Qualitative inquiry in the history of psychology. *Qual. Psychol.* 1, 4–16. doi: 10.1037/qup0000007

Wiggins, S., Gordon-Finlayson, A., Becker, S., and Sullivan, C. (2016). Qualitative undergraduate project supervision in psychology: current practices and support needs of supervisors across north east England and Scotland. *Qual. Res. Psychol.* 13, 1–19. doi: 10.1080/14780887.2015.1075641

Scientific integrity in research methods

Jordan R. Schoenherr *

Department of Psychology, Carleton University, Ottawa, ON, Canada

INTRODUCTION

Recent cases of research misconduct have prompted psychologists to suggest that there is too much vulnerability in the research process (e.g., Simmons et al., 2011; Pashler and Harris, 2012). Regardless of whether this is the case, ensuring the integrity of a discipline requires a clear understanding of what conventions and norms define the research process. In what follows, I consider the research integrity curriculum of North American psychology. In particular, I will claim that a major impediment to ensuring responsible research practices is an underspecified and understudied curriculum.

DEFINING CURRICULUM

A general distinction made in the education literature is that between the explicit curriculum and implicit curriculum (e.g., Posner, 1992; Palomba and Banta, 1999). The *explicit curriculum* (EC) consists of the information in courses, textbooks, and workshops that are formally provided to learners. The EC contains the core concepts, norms, and values of an academic discipline. In that the content of the curriculum is clearly specified, the EC requires that learners achieve mastery of these theories, methods, and analytic skills. The *implicit curriculum* (IC) consists of the information that learners acquire throughout their studies that is not included in the explicit curriculum. The IC contains information that qualifies the formal curriculum such as exceptions to rules, as well as tacit knowledge or craft skills (e.g., Polanyi, 1958; Latour and Woolgar, 1979; Charlesworth et al., 1989). Craft skills can include how to select the "right" research question, how to effectively use experimental, analytic, and graphical software, and how to frame publications for acceptance. In that the IC is *ad hoc*, it can be considered an apprenticeship that relies on the specific knowledge and attention given to the learner by supervisors, mentors, and instructors.

THE STANDARDS OF PSYCHOLOGY

Key issues in scientific integrity have been outlined by governmental and non-governmental organizations in North America. In the United States, governmental standards have been provided by the Office of Research Integrity in terms of the responsible conduct of research (Steneck, 2006). In Canada, major granting organizations have provided general standards that are conditions of receiving research funds (Tri-council, 2006). University policies often extend these general standards but are highly variable in terms of their content (e.g., Greene et al., 1985; Lind, 2005; Schoenherr and Williams-Jones, 2011). For instance, while data falsification is universally agreed upon as deviant behavior, publication practices are not addressed to the same extent. Explicit and implicit curricula are left to address these concerns.

Professional organizations such as the American Psychological Association (APA) provide standards for responsible research practices to their members. In addition to the five general principles of conduct (beneficence and non-maleficence, fidelity and responsibility, integrity, justice, and respect for people's rights and dignity), the APA also identifies 10 guidelines related to scientific integrity (Section 8.10a to 8.15; American Psychological Association, 2002/2010). These APA guidelines address the reporting of research results (fabrication and error correction), plagiarism, publication credit (inclusion criteria, contributions, and student credit), duplicate publication, data sharing (post-publication and limitations), and roles and responsibilities of reviewers. Whether, and how, these norms are presented within the curriculum is an open question that I will briefly consider below.

UNDERGRADUATE CURRICULUM

The dearth of research on issues related to scientific integrity in psychological science can be contrasted with repeated reviews of the psychology curriculum more generally (Henry, 1938; Sanford and Fleishman, 1950; Daniel et al., 1965; Kulik, 1973; Lux and Daniel, 1978; Scheirer and Rogers, 1985; Cooney and Griffith, 1994; Perlman and McCann, 2005). For instance, Perlman and McCann (1999a) sampled the course catalogs of 400 institutions (evenly split among doctoral, comprehensive, baccalaureate, and 2-year colleges) from 1961 to 1997 to examine the courses offered by psychology departments. Courses that are provided across institutions provide insight into the family resemblance structure of the explicit scientific integrity curriculum within psychological science. If scientific integrity is viewed in terms of appropriate conduct through the planning, implementation, analysis, interpretation, and communication of research findings, identifying courses that likely contain this information provides a test for how scientific integrity is presented (Ellis, 1992). Thus, courses addressing experimental design, research methods, statistics, tests and measurement, as well as field experience represent an important starting point for queries into the scientific integrity curriculum (see **Table 1**).

Perlman and McCann's (1999a) study can be interpreted as evidence that scientific integrity-related courses were recurrent features of the psychology curriculum in most academic institutions. However, whether we consider all academic institutions that were sampled or solely doctoral institutions, it is clear that if these courses address scientific integrity issues then these issues might only be addressed in an inconsistent manner. In a follow-up study conducted by Perlman and McCann (2005), they also observed that courses that provide students with research experience were not obligatory and that there was considerable interdepartmental variability in terms of when these courses were offered. This underscores the importance of considering degree requirements.

A stronger test of the scientific integrity curriculum is to consider courses that are included in students' degree requirements. In a companion analysis of the structure of degrees in psychological science, Perlman and McCann (1999b) consider what courses were listed as degree requirements in 500 institutions. The majority of institutions listed capstone (i.e., courses that require integrating knowledge of theory and methods; 63%) and statistics (58%) as requirements while research methods courses (40%) and experimental psychology (38%) were required to a lesser extent. Other courses related to scientific integrity such as psychometrics (9%) and experimental design (7%) were degree requirements in the minority of institutions. In a comparable manner to their study of courses offered by institutions (Perlman and McCann, 1999a), Perlman and McCann (1999b) also note that degree requirements differed based on the type of institutions. For instance, whereas the majority of doctoral institutions required statistics courses (65%), comprehensive (59%) and baccalaureate institutions (49%) did so to a lesser extent. The variability in courses offered (Perlman and McCann, 1999a) and the extent to which they constitute degree requirements (Perlman and McCann, 1999b) suggest that the EC might only weakly addresses issues of scientific integrity.

Further support for curriculum variability is evidenced in the contents of research methods textbooks. Textbooks are a means to present ideal disciplinary standards in terms of core theories and evidence (e.g., Ash, 1983; Weiten and Wight, 1992; Zechmeister and Zechmeister, 2000). While research methods textbooks typically discuss issues of design (e.g., distinguishing between dependent and independent variables, participant selection, within-, or between-subjects design), scientific integrity

TABLE 1 | Percentage of institutions sampled by Perlman and McCann (1999a) that offer undergraduate curriculum related to scientific integrity either for all institutions sampled (All) and the subsample of doctoral institutions (DU).

Course related to scientific integrity	Year of university samples								DU Δ (97–75)
	1961		1969		1975		1997		
	All	DU	All	DU	All	DU	All	DU	
Tests and measurement	48	84	48	81	51	**79**	51	**68**	**−11**
Statistics	36	74	43	75	46	**70**	48	**72**	**+2**
Experimental	45	84	45	73	39	**50**	44	**57**	**+7**
Field experience	–	–	–	–	30	**32**	30	**44**	**+12**
Research methods	–	–	–	–	42	**34**	42	**58**	**+24**

Bold numbers reflect values used to obtain difference score. Difference score reflects change in course requirements from 1975 to 1997.

issues (e.g., conflict of interests, data fabrication, publication practices) might not be included. Importantly, while research *ethics* is a near ubiquitous feature in research methods textbooks, these issues are restricted to the treatment of human and non-human participants. Although, research methods textbooks have begun to discuss misconduct, the minority mention the APA guidelines that address scientific integrity. For instance, an examination of a sample of research methods textbooks used at the author's institution (e.g., Smith and Davis, 2003; Shaughnessy et al., 2006; McBurney and White, 2010; Cozby and Rawn, 2012; Gravetter and Forzano, 2012; Leary, 2012) revealed that no textbook included all of these guidelines and that there is considerable variability in how many are discussed. While fabrication, error correction, and plagiarism were the most common forms of misconduct discussed, various aspects of publication credit and data sharing were addressed to a lesser extent. As with the EC, research methods textbooks used to support these courses do not appear to address the issues of scientific integrity in a comprehensive or consistent manner. This leaves the responsibility for scientific integrity education in the hands of individual instructors, supervisors, and mentors.

EARLY EXPERIENCES AND GRADUATE MENTORSHIP

It might be argued that many undergraduates neither necessarily seek, nor are considered for, graduate studies. Consequently, evaluation of the undergraduate psychology curriculum might not be the best approach to examining scientific integrity issues. For instance, while graduate and post-graduate researchers are concerned with research and publication, undergraduates need not be instructed in the specific practices required to conduct research. This argument reflects specious reasoning. First, higher education is directed toward understanding a research area. Researchers must understand the basic theories, experimental methods, and analytic procedures they use directly or indirectly (e.g., Schoenherr and Hamstra, 2015). This will minimally make students better consumers of scientific knowledge. Second, both socialization and expertise development require repeated use of social conventions, declarative knowledge, and technical skill. Given graduate students' first experience with responsible research practices occurs within the undergraduate curriculum, setting an early precedent is necessary. Moreover, as Lovitts (2007) has noted in the context of the doctoral dissertation, explicit conventions are often absent or not communicated to students. Michell (1997) goes further to claim that "many psychological researchers are ignorant with respect to the methods they use… the ignorance I refer to is about the *logic* of methodological practices," (p. 356). If true, this suggests graduate school is not providing adequate instruction to develop these competencies.

A likely cause is revealed when we reflect on the experiences of graduate students. Much graduate work is based on self-directed learning. While courses are offered in advanced statistical techniques (e.g., multidimensional scaling, hierarchical linear modeling, factor analysis), other aspects of research methods are alluded to in research articles or left to supervisors and mentors to explicate. As research articles are a genre and limited in the extent to which they can discuss the research process, much of the scientific integrity curriculum is necessarily implicit. It is therefore likely to vary depending on the competency and experience of faculty members, reinforcing the importance of mentorship in education in general (e.g., Bird, 2001; Paglis et al., 2006; Anderson et al., 2007) and psychology in particular (e.g., Cronan-Hillix et al., 1986; Clark et al., 2000; Forehand, 2008). Concerns over the sufficiency of this form of apprenticeship must be addressed.

Once apprenticeship is recognized as a central feature of graduate studies, the extent to which psychologists share beliefs about scientific integrity becomes a central concern. However, psychologists have been found to disagree over the priority of APA standards (Seitz and O'Neill, 1996; Hadjistavropoulos et al., 2002) and are inconsistent in their application (Williams et al., 2012). Similar results have been observed for issues of scientific integrity. Riordan et al. (1988) examined psychologists' perceptions of plagiarism and fabrication. They note that while fabrication was viewed as more detrimental to a researcher's career, psychologists believed that university action was more justified in cases of plagiarism. More recently, John et al. (2012) assessed the prevalence and perceptions of questionable research practices by psychologists. They found that the manipulation of results in an unplanned and unreported manner was a reasonably common practice while also being judged to be dishonest by psychologists.

CONCLUSIONS

Psychology is no more susceptible to disagreement over its norms than any other science (Ioannidis, 2005; De Vries et al., 2006). Consequently, variability of the undergraduate and graduate curricula suggests that a more *explicit* treatment of scientific integrity issues should be pursued. Despite the possibility that undergraduate statistics and research methods courses might address some of these issues in a general manner, other topics are not likely to be addressed. This appears to be reflected in the variable content of research methods textbooks. If departments are unclear as to whether this is the case, tools such as curriculum matrices (e.g., Levy et al., 1999) can be used to formally evaluate the features of their curriculum. Curriculum matrices require that faculty members identify core topics that should be covered within a curriculum and assess which courses address this information. When course information is plotted on such a grid, gaps are revealed and can then be addressed. In conjunction with the standards of professional organizations, formal policies, and guidelines can also be developed to ensure greater consistency.

FUNDING

REFERENCES

American Psychological Association. (2002/2010). Ethical principles of psychologists and code of conduct. *Am. Psychol.* 57, 1060–1073.

Anderson, M. S., Horn, A. S., Risbey, K. R., Ronning, E. A., De Vries, R., and Martinson, B. C. (2007). What do mentoring and training in the responsible conduct of research have to do with scientists' misbehavior?: findings from a national survey of NIH-funded scientists. *Acad. Med.* 82, 853–860. doi: 10.1097/ACM.0b013e31812f764c

Ash, M. G. (1983). "The self-presentation of a discipline: history of psychology in the United States between pedagogy and scholarship," in *Functions and Uses of Disciplinary Histories,* eds L. Graham, W. Lepenies, and P. Weingart (Dordrecht: Reidel), 143–189.

Bird, S. J. (2001). Mentors, advisors and supervisors: their role in teaching responsible research conduct. *Sci. Eng. Ethics* 7, 455–468. doi: 10.1007/s11948-001-0002-1

Charlesworth, M., Farrall, L., Stokes, T., and Turnbull, D. (1989). *Life Among The Scientists.* Melbourne: Oxford University Press.

Clark, R. A., Harden, S. L., and Johnson, W. B. (2000). Mentor relationships in clinical psychology doctoral training: results of a national survey. *Teach. Psychol.* 27, 262–268. doi: 10.1207/S15328023TOP2704_04

Cooney, B. R., and Griffith, D. M. (1994). *The 1992.1993 Undergraduate Department Survey.* Washington, DC: American Psychological Association.

Cozby, P. C., and Rawn, C. D. (2012). *Methods in Behavioural Research, Canadian Edition.* McGraw-Hill Ryerson.

Cronan-Hillix, T., Gensheimer, L. K., Cronan-Hillix, W. A., and Davidson, W. S. (1986). Students' views of mentors in psychology graduate training. *Teach. Psychol.* 13, 123–127. doi: 10.1207/s15328023top1303_5

Daniel, R. S., Dunham, P. J., and Morris, C. J. Jr. (1965). Undergraduate courses in psychology; 14 years later. *Psychol. Rec.* 15, 25–31.

De Vries, R., Andersen, M. S., and Martinson, B. C. (2006). Normal misbehaviour: scientists talk about the ethics of research. *J. Empir. Res. Hum. Res. Ethics* 1, 43–50. doi: 10.1525/jer.2006.1.1.43

Ellis, H. C. (1992). Graduate education in psychology: past, present, and future. *Am. Psychol.* 47, 570–576. doi: 10.1037/0003-066X.47.4.570

Forehand, R. L. (2008). The art and science of mentoring in psychology: a necessary practice to ensure our future. *Am. Psychol.* 63, 744–755. doi: 10.1037/0003-066X.63.8.744

Gravetter, F. J., and Forzano, L. B. (2012). *Research Methods for the Behavioral Sciences, 4th Edn.* Belmont, CA: Wadsworth Cengage Learning.

Greene, P. J., Durch, J. S., Horwitz, W., and Hooper, V. (1985). Policies for responding to allegations of fraud in research. *Minerva* 23, 203–215. doi: 10.1007/BF01099942

Hadjistavropoulos, T., Malloy, D. C., Sharpe, D., Green, S. M., and Fuchs-Lacelle, S. (2002). The relative importance of the ethical principles adopted by the American Psychological Association. *Can. Psychol.* 43, 254–259. doi: 10.1037/h0086921

Henry, E. R. (1938). A survey of courses in psychology offered by undergraduate colleges of liberal arts. *Psychol. Bullet.* 35, 430–435. doi: 10.1037/h0059009

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Kulik, J. A. (1973). *Undergraduate Education in Psychology.* Washington, DC: American Psychological Association.

Latour, B., and Woolgar, S. (1979). *Laboratory Life: The Construction of Scientific Facts.* Princeton: Princeton University Press.

Leary, M. R. (2012). *Introduction to Behavioral Research Methods, 6th Edn.* New York, NY: Pearson.

Levy, J., Burton, G., Mickler, S., and Vigorito, M. (1999). A curriculum matrix for psychology program review. *Teach. Psychol.* 26, 291–294.

Lind, R. A. (2005). Evaluating research misconduct policies at major research universities: a pilot study. *Account. Res.* 12, 241–262. doi: 10.1080/08989620500217560

Lovitts, B. E. (ed.). (2007). "The psychology dissertation" in *Making the Implicit Explicit: Faculty's Performance Expectations for the Dissertation* (Sterling, TX: Stlyus), 247–270.

Lux, D. F., and Daniel, R. S. (1978). Which courses are most frequently listed by psychology departments? *Teach. Psychol.* 5, 13–16.

McBurney, D. H., and White, T. L. (2010). *Research Methods, 8th Edn.* Belmont, CA: Wadsworth Cengage Learning.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *Br. J. Psychol.* 88, 355–383. doi: 10.1111/j.2044-8295.1997.tb02641.x

Paglis, L. L., Green, S. G., and Bauer, T. N. (2006). Does adviser mentoring add value? A longitudinal study of mentoring in doctoral student outcomes. *Res. High. Educ.* 47, 451–476. doi: 10.1007/s11162-005-9003-2

Palomba, C., and Banta, T. (1999). *Curriculum Assessment Essentials: Planning, Implementing & Improving Assessment in Higher Education.* San Francisco, CA: Jossey-Bass.

Pashler, H., and Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* 7, 531–536. doi: 10.1177/1745691612463401

Perlman, B., and McCann, L. I. (1999a). The most frequently listed courses in the undergraduate psychology curriculum. *Teach. Psychol.* 26, 177–182.

Perlman, B., and McCann, L. I. (1999b). The structure of the psychology undergraduate curriculum. *Teach. Psychol.* 26, 171–176.

Perlman, B., and McCann, L. I. (2005). Undergraduate research experiences in Psychology: a national study of courses and curricula. *Teach. Psychol.* 32, 5–14. doi: 10.1207/s15328023top3201_2

Polanyi, M. (1958). *Personal Knowledge: Toward a Post-Critical Philosophy.* Chicago, IL: University of Chicago Press.

Posner, G. J. (1992). *Analyzing the Curriculum.* New York, NY: McGraw-Hill, Inc.

Riordan, C. A., Marlin, N. A., and Gidwani, C. (1988). Accounts offered for unethical Research practices: effects on the evaluations of acts and actors. *J. Soc. Psychol.* 128, 495–505. doi: 10.1080/00224545.1988.9713769

Sanford, F. H., and Fleishman, E. A. (1950). A survey of undergraduate psychology courses in American colleges and universities. *Am. Psychol.* 5, 33–37.

Scheirer, C. J., and Rogers, A. M. (1985). *The Undergraduate Psychology Curriculum: 1984.* Washington, DC: American Psychological Association.

Schoenherr, J. R., and Hamstra, S. J. (2015). Psychometrics and its discontents: a historical perspective on the discourse of the measurement tradition. *Advan. Health Sci. Educ.* doi: 10.1007/s10459-015-9623-z. [Epub ahead of print].

Schoenherr, J., and Williams-Jones, B. (2011). Research integrity/misconduct policies of Canadian universities. *Can. J. High. Educ.* 41, 1–17.

Seitz, J., and O'Neill, P. (1996). Ethical decision-making and the code of ethics of the Canadian Psychological Association. *Can. Psychol.* 37, 21–30. doi: 10.1037/0708-5591.37.1.23

Shaughnessy, J. J., Zechmeister, E. B., and Zechmeister, J. S. (2006). *Research Methods in Psychology, 7th Edn.* New York, NY: McGrawHill.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Smith, R. A., and Davis, S. F. (2003). *The Psychologist as Detective, 3rd Edn.* Upper Saddle River, NJ: Pearson Prentice Hall.

Steneck, N. (2006). Fostering integrity in research: definitions, current knowledge, and future directions. *Sci. Eng. Ethics* 12, 53–74. doi: 10.1007/s11948-006-0006-y

Tri-council. (2006). *Tri-Council Policy Statement: Integrity in Research and Scholarship.* Available online at: http://www.nserc.gc.ca/professors_e.asp?nav=profnav&lbi=p9

Weiten, W., and Wight, R. D. (1992). "Portraits of a discipline: an examination of introductory psychology textbooks in America," in *Teaching Psychology*

in America: A History, eds A. E. Puente, J. R. Matthews, and C. L. Brewer (Washington, DC: American Psychological Association), 453–504.

Williams, J., Hadjistavropoulos, T., Malloy, D. C., Gagnon, M., Sharpe, D., and Fuchs-Lacelle, S. (2012). A mixed methods investigation of the effects of ranking ethical principles on decision making: implications for the canadian code of ethics for psychologists. *Can. Psychol.* 53, 204–216. doi: 10.1037/a00 27624

Zechmeister, J. S., and Zechmeister, E. B. (2000). Introductory textbooks and psychology's core concepts. *Teach. Psychol.* 27, 6–11. doi: 10.1207/S15328023TOP2701_1

Advantages of publishing in Frontiers

OPEN ACCESS

Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW

Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION

Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS

No limit to article distribution and re-use

TRANSPARENT

Editors and reviewers acknowledged by name on published articles

SUPPORT

By our Swiss-based editorial team

IMPACT METRICS

Advanced metrics track your article's impact

GLOBAL SPREAD

5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK

Our network increases readership for your article