frontiers | Research Topics

Human-centered AI: Crowd computing

Edited by
Jie Yang, Alessandro Bozzon, Ujwal Gadiraju
and Matthew Lease

Published in
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-2502-9
DOI 10.3389/978-2-8325-2502-9

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: frontiersin.org/about/contact

Human-centered AI: Crowd computing

Topic editors
Jie Yang — Delft University of Technology, Netherlands
Alessandro Bozzon — Delft University of Technology, Netherlands
Ujwal Gadiraju — Delft University of Technology, Netherlands
Matthew Lease — The University of Texas at Austin, United States

Citation
Yang, J., Bozzon, A., Gadiraju, U., Lease, M., eds. (2023). *Human-centered AI: Crowd computing*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2502-9

Table of contents

frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 15 March 2023
DOI 10.3389/frai.2023.1161006

Editorial: Human-centered AI: Crowd computing

Jie Yang[1]*, Alessandro Bozzon[2], Ujwal Gadiraju[1] and Matthew Lease[3]

[1]Web Information Systems, Delft University of Technology, Delft, Netherlands, [2]Knowledge and Intelligence Design, Delft University of Technology, Delft, Netherlands, [3]School of Information, The University of Texas at Austin, Austin, TX, United States

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Claudia Wagner,
GESIS Leibniz Institute for the Social Sciences, Germany

*CORRESPONDENCE
Jie Yang
✉ j.yang-3@tudelft.nl

SPECIALTY SECTION
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 07 February 2023
ACCEPTED 27 February 2023
PUBLISHED 15 March 2023

CITATION
Yang J, Bozzon A, Gadiraju U and Lease M (2023) Editorial: Human-centered AI: Crowd computing. *Front. Artif. Intell.* 6:1161006.
doi: 10.3389/frai.2023.1161006

KEYWORDS

human-centered AI, human-in-the-loop AI, human-AI interaction, human computation, crowdsourcing

Editorial on the Research Topic
Human-centered AI: Crowd computing

1. Introduction

Human computation (HCOMP) and crowdsourcing (Law and von Ahn, 2011; Quinn and Bederson, 2011; Kittur et al., 2013; Lease and Alonso, 2018) have been instrumental to advances seen in artificial intelligence (AI) and machine learning (ML) over the past 15+ years. AI/ML has an insatiable hunger for human labeled training to supervise models, with training data scale playing a significant (if not dominant) role in driving the predictive performance of models (Halevy et al., 2009). The centrality of such human-labeled data to the success and continuing advancement of AI/ML is thus at the heart of today's *data-centric AI* movement (Mazumder et al., 2022). Moreover, recent calls for *data excellence* (Aroyo et al., 2022) reflect growing recognition that AI/ML data scale alone does not suffice. The quality of human labeled data also plays a tremendous role in AI/ML success, and ignoring this can be perilous to deployed AI/ML systems (Sambasivan et al., 2021), as prominent, public failures have shown.

HOMP and crowdsourcing have also enabled hybrid, human-in-the-loop, crowd-powered computing (Demartini et al., 2017). When state-of-the-art AI/ML cannot provide sufficient capabilities or predictive performance to meet practical needs for real-world deployment, hybrid systems utilize HCOMP at run-time to deliver last-mile capabilities where AI/ML fall short (Gadiraju and Yang, 2020). This has enabled a new class of innovative and more capable applications, systems, and companies to be built (Barr and Cabrera, 2006). While work in HCOMP is centuries old (Grier, 2013), access to an increasingly Internet-connected and well-educated world population led to the advent of crowdsourcing (Howe, 2006). This has allowed AI/ML systems to call on human help at run-time as "Human Processing Units (HPUs)"(Davis et al., 2010), "Remote Person Calls (RPCs)" (Bederson and Quinn, 2011), and "the Human API" (Irani and Silberman, 2013).

Across both data labeling and run-time HCOMP, crowdsourcing has enabled AI/ML builders to tap into the "wisdom of the crowd" (Surowiecki, 2005) and harness *collective intelligence* from large groups of people. As AI/ML systems have grown both more powerful and ubiquitous, appreciation of their capabilities has also been tempered by concerns of prevalence and propagation of biases, lack of robustness, fairness, and transparency as well as ethical and societal implications. At the same time, crowdsourced access to a global, diverse set of contributors provides an incredible avenue to boost inclusivity and fairness in both AI/ML labeled datasets and hybrid, human-in-the-loop systems. However, important questions remain about the roles and treatment of AI/ML data workers, and the extent to which AI/ML advances are creating new economic opportunities for human workers (Paritosh et al., 2011) or exploiting hidden human labor (Bederson and Quinn, 2011; Fort et al., 2011; Irani and Silberman, 2013; Lease and Alonso, 2018; Gray and Suri, 2019). This has prompted the development of ethical principles for crowd work (Graham et al., 2020) and calls for *responsible sourcing* of AI/ML data (Partnership on AI, 2021).

As the above discussion reflects, HCOMP and crowdsourcing reflects a rich amalgamation of interdisciplinary research. In particular, the confluence of two key research communities—AI/ML and human-computer interaction (HCI)—has been central to founding and advancing HCOMP and crowdsourcing. Beyond this, related work draws upon a wide and rich body of diverse areas, including (but not limited to) computational social science, digital humanities, economics, ethics, law / policy / regulation, and social computing. More broadly, the HCOMP and crowdsourcing community promotes the exchange of advances in the state-of-the-art and best practices not only among researchers but also engineers and practitioners, to encourage dialogue across disciplines and communities of practice.

2. Call for papers: Aim and scope

Our organization of this Frontiers *Research Topic* called for new and high-impact contributions in HCOMP and crowdsourcing. We especially encouraged work that generates new insights into the collaboration and interaction between humans and AI, enlarging understanding about hybrid human-in-the-loop and algorithm-in-the-loop systems (Green and Chen, 2020). This includes human-AI interaction, algorithmic and interface techniques for augmenting human abilities to AI systems. It also spans issues that affect how humans collaborate and interact with AI systems such as bias, interpretability, usability, and trustworthiness. We welcomed both system-centered and human-centered approaches to human+AI systems, considering humans as users and stakeholders, or as active contributors and an integral part of the system.

Our call for papers invited submissions relevant to theory, studies, tools and/or applications that present novel, interesting, impactful interactions between people and computational systems. These cover a broad range of scenarios across human computation, wisdom of the crowds, crowdsourcing, and people-centric AI methods, systems and applications.

The scope of the Research Topics included the following themes:

- Crowdsourcing applications and techniques.
- Techniques that enable and enhance human-in-the-loop systems, making them more efficient, accurate, and human-friendly.
- Studies about how people perform tasks individually, in groups, or as a crowd.
- Approaches to make crowd science FAIR (Findable, Accessible, Interoperable, Reproducible) and studies assessing and commenting on the FAIRness of human computation and crowdsourcing practice.
- Studies into fairness, accountability, transparency, ethics, and policy implications for crowdsourcing and human computation.
- Methods that use human computation and crowdsourcing to build people-centric AI systems and applications, including topics such as reliability, interpretability, usability, and trustworthiness.
- Studies into the reliability and other quality aspects of human-annotated and -curated datasets, especially for AI systems.
- Studies about how people and intelligent systems interact and collaborate with each other and studies revealing the influences and impact of intelligent systems on society.
- Crowdsourcing studies into the socio-technical aspects of AI systems: privacy, bias, and trust.

3. Partnership with AAAI HCOMP

For over a decade, the premier venue for disseminating the latest research findings on HCOMP and crowdsourcing has been the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Human Computation and Crowdsourcing (AAAI HCOMP).[1] Early HCOMP workshops at KDD and AAAI conferences (2009-2012) led to the genesis of the AAAI HCOMP conference in 2013. To further strengthen this Frontiers *Research Topic*, we partnered with AAAI HCOMP to invite submissions; papers accepted to HCOMP 2021 were offered a streamlined process for publication in this topic (e.g., maintaining the same reviewers when possible). We accepted four such submissions that extend earlier HCOMP 2021 papers (Samiotis et al., 2021; Welty et al., 2021; Yamanaka, 2021; Yasmin et al., 2021).

4. Managing conflicts-of-interest (COI)

"The Frontiers review system is designed to guarantee the most transparent and objective editorial and review process, and because the handling editor's and reviewers' names are made public upon the publication of articles, conflicts of interest will be widely apparent" (Frontiers, 2023). For this Frontiers *Research Topic*, two submissions from topic editors were routed by Frontiers staff to other editors not otherwise associated with this *Research Topic* and had no COI with the topic editors. Both submissions were ultimately accepted (Pradhan et al.; Samiotis et al.), after which the identity of each handling editor became publicly available. We thank these additional editors for their contributions to this *Research Topic*.

5. Research Topic contributions

A total of nine articles were accepted, contributing studies into factors of human computation and crowdsourcing, to their

[1] https://humancomputation.com/

applications to human-AI collaborative systems and large-scale behavioral studies. In the following, we very briefly summarize these works.

5.1. Quality in crowdsourced data annotation

Annotation quality is often a key concern in crowdsourced labeling. Pradhan et al. introduce a three-stage FIND-RESOLVE-LABEL workflow to reduce ambiguity in annotation task instructions. Their workflow allows workers to provide feedback on ambiguous task instructions to a requester. Another aspect of annotation quality is worker disagreement, for which a number of methods have been developed. Drawing from the observation that the effectiveness of annotation depends on the level of noise in the data, Uma et al. investigate the use of temperature scaling to estimate noise. Yasmin et al. investigates the effect of different forms of input elicitation to improve the quality of inferred labels in image classification, suggesting that more accurate results can be achieved when labels and self-reported confidence are used as features for classifiers.

5.2. Human-centered computation and interaction in AI

Tocchetti et al. study the effect of gamified activities to improve crowds' understanding of black-box models, addressing the intelligibility issue of explainable AI. They consider gamified activities to educate crowds by AI researchers. Yamanaka investigates the effectiveness of crowdsourcing for validating user performance models, especially the error-rate prediction model in target pointing tasks, which requires data from many repetitive experiments by participants for each task condition to measure the central tendency of the error rate. Welty et al. studies crowd knowledge creation for curating class-level knowledge graphs. Their three-tier crowd approach to elicit class-level attributes addresses the label sparsity problem faced by AI/ML systems.

5.3. Human factors in human computation

Vinella, Hu et al. focuses the effect of human agency in team formation on team performance. They found that in open collaboration scenarios, e.g., hackathon, teams formed by workers themselves are more competitive, compared to those formed by algorithms. Samiotis et al. explore the possession of musical skills in the worker population. Their study shows that untrained workers possess high perception skills that can be useful in many music annotation tasks. Vinella, Odo et al. study the effect of personality on task performance by ad-hoc teams composed of strangers, especially in solving critical tasks that are often time-bounded and high-stress, e.g., incident response. Their results identify personality traits that affect team performance and in addition to that, relevant communication patterns used by winning teams.

Author contributions

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

Acknowledgments

We would like to thank the AAAI HCOMP Steering Committee, as well as the organizers of the 2021 conference https://www.humancomputation.com/2021/, for their partnership with our Research Topic on Human-Centered AI: Crowd Computing. ML was supported in part by Good Systems https://goodsystems.utexas.edu, a UT Austin Grand Challenge to develop responsible AI technologies. Any opinions expressed in this article are entirely our own.

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aroyo, L., Lease, M., Paritosh, P., and Schaekermann, M. (2022). Data excellence for AI: why should you care? *Interactions* 29, 66–69. doi: 10.1145/3517337

Barr, J., and Cabrera, L. F. (2006). AI gets a brain: new technology allows software to tap real human intelligence. *Queue* 4, 24–29. doi: 10.1145/1142055.1142067

Bederson, B. B., and Quinn, A. J. (2011). "Web workers unite! addressing challenges of online laborers," in *CHI Workshop on Crowdsourcing and Human Computation* (Vancouver, BC: ACM).

Davis, J., Arderiu, J., Lin, H., Nevins, Z., Schuon, S., Gallo, O., et al. (2010). "The HPU," in *Computer Vision and Pattern Recognition Workshops (CVPRW)* (San Francisco, CA), 9–16.

Demartini, G., Difallah, D. E., Gadiraju, U., and Catasta, M. (2017). An introduction to hybrid human-machine information systems. *Found. Trends® Web Sci.* 7, 1–87. doi: 10.1561/1800000025

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon mechanical turk: gold mine or coal mine? *Comput. Linguist.* 37, 413–420. doi: 10.1162/COLI_a_00057

Frontiers (2023). *Policies and Publication Ethics*. Available online at: https://www.frontiersin.org/guidelines/policies-and-publication-ethics/.

Gadiraju, U., and Yang, J. (2020). 'What can crowd computing do for the next generation of ai systems?," in *2020 Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation* (CEUR), 7–13.

Graham, M., Woodcock, J., Heeks, R., Mungai, P., Van Belle, J.-P., du Toit, D., et al. (2020). The fairwork foundation: strategies for improving platform work in a global context. *Geoforum* 112, 100–103. doi: 10.1016/j.geoforum.2020.01.023

Gray, M. L., and Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley From Building a New Global Underclass*. Eamon Dolan Books.

Green, B., and Chen, Y. (2020). "Algorithm-in-the-loop decision making," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, NY), Vol. 34, 13663–13664.

Grier, D. A. (2013). *When Computers Were Human*. Princeton University Press.

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intellig. Syst.* 24, 8–12. doi: 10.1109/MIS.2009.36

Howe, J. (2006). The rise of crowdsourcing. *Wired Magaz.* 14, 1–4.

Irani, L., and Silberman, M. (2013). "Turkopticon: interrupting worker invisibility in amazon mechanical turk," in *Proceeding of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Paris).

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., et al. (2013). "The future of crowd work," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)* (San Antonio, TX), 1301–1318.

Law, E., and von Ahn, L. (2011). Human computation. *Synth. Lectur. Artif. Intellig. Mach. Learn.* 5, 1–121. doi: 10.1007/978-3-031-01555-7

Lease, M., and Alonso, O. (2018). "Crowdsourcing and human computation: introduction," in *Encyclopedia of Social Network Analysis and Mining*, eds R. Alhajj and J. Rokne (New York, NY: Springer), 499–510.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Rojas, W. G., Diamos, S., et al. (2022). Dataperf: benchmarks for data-centric AI development. *arXiv preprint* arXiv:2207.10062.

Paritosh, P., Ipeirotis, P., Cooper, M., and Suri, S. (2011). "The computer is the new sewing machine: benefits and perils of crowdsourcing," in *Proceedings of the 20th International Conference Companion on World Wide Web* (Hyderabad: ACM), 325–326.

Partnership on AI (2021). *Responsible Sourcing Across the Data Supply Line*. Available online at: https://partnershiponai.org/workstream/responsible-sourcing/.

Quinn, A. J., and Bederson, B. B. (2011). "Human computation: a survey and taxonomy of a growing field," in *2011 Annual ACM SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC), 1403–1412.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. (2021). ""Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Samiotis, I. P., Qiu, S., Lofi, C., Yang, J., Gadiraju, U., and Bozzon, A. (2021). "Exploring the music perception skills of crowd workers," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9, 108–119.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Welty, C., Aroyo, L., Korn, F., McCarthy, S. M., and Zhao, S. (2021). "Rapid instance-level knowledge acquisition for google maps from class-level common sense," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, 143–154.

Yamanaka, S. (2021). "Utility of crowdsourced user experiments for measuring the central tendency of user performance to evaluate error-rate models on guis," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9, 155–165.

Yasmin, R., Grassel, J. T., Hassan, M. M., Fuentes, O., and Escobedo, A. R. (2021). "Enhancing image classification capabilities of crowdsourcing-based methods through expanded input elicitation," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9, 166–178.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/frai.2022.830299

Addressing Label Sparsity With Class-Level Common Sense for Google Maps

Chris Welty, Lora Aroyo, Flip Korn, Sara M. McCarthy and Shubin Zhao*

Google Research, New York, NY, United States

Successful knowledge graphs (KGs) solved the historical knowledge acquisition bottleneck by supplanting the previous expert focus with a simple, crowd-friendly one: KG nodes represent popular people, places, organizations, etc., and the graph arcs represent common sense relations like affiliations, locations, etc. Techniques for more general, categorical, KG curation do not seem to have made the same transition: the KG research community is still largely focused on logic-based methods that belie the common-sense characteristics of successful KGs. In this paper, we propose a simple yet novel three-tier crowd approach to acquiring *class-level attributes* that represent broad common sense associations between categories, and can be used with the classic knowledge-base default & override technique, to address the early *label sparsity problem* faced by machine learning systems for problems that lack data for training. We demonstrate the effectiveness of our acquisition and reasoning approach on a pair of very real industrial-scale problems: how to augment an existing KG of places and offerings (e.g. stores and products, restaurants and dishes) with associations between them indicating the availability of the offerings at those places. Label sparsity is a general problem, and not specific to these use cases, that prevents modern AI and machine learning techniques from applying to many applications for which labeled data is not readily available. As a result, the study of how to acquire the knowledge and data needed for AI to work is as much a problem today as it was in the 1970s and 80s during the advent of expert systems. Our approach was a critical part of enabling a worldwide *local search* capability on Google Maps, with which users can find products and dishes that are available in most places on earth.

Keywords: map, knowledge graph, crowdsourcing, class-level attributes, common sense, knowledge acquisition

OPEN ACCESS

Edited by:
Jie Yang,
Delft University of Technology,
Netherlands

Reviewed by:
Lorenzo Corti,
Politecnico di Milano, Italy
Agathe Balayn,
Delft University of Technology,
Netherlands

****Correspondence:***
Chris Welty
cawelty@gmail.com

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

Received: *07 December 2021*
Accepted: *25 January 2022*
Published: *16 March 2022*

Citation:
Welty C, Aroyo L, Korn F, McCarthy SM and Zhao S (2022) Addressing Label Sparsity With Class-Level Common Sense for Google Maps.
Front. Artif. Intell. 5:830299.
doi: 10.3389/frai.2022.830299

1. INTRODUCTION

From the outset, knowledge graphs (KGs) have prominently used crowdsourcing for knowledge acquisition, both from the perspective of scaling out graph creation and long-term maintenance, solving the historical knowledge acquisition bottleneck by revisiting the expert systems assumption that knowledge should be acquired from experts. As a result, popular KGs like Freebase (Bollacker et al., 2008)—now

Google's Knowledge Graph—and WikiData (Vrandečić and Krötzsch, 2014) are composed primarily of popular "common sense" entities and relations in the world that people are exposed to regularly and that can be acquired from and validated by the crowd.

Similarly, today Google Maps overlays data on maps about the different places or establishments (stores, restaurants, hospitals, etc.) worldwide, and crowdsourcing plays a central role in the acquisition and maintenance of this information, as discussed in Lagos et al. (2020). Users contribute opening hours, locations, reviews, etc., as well as categorical information about places such as whether it is a supermarket, department store, etc., which makes KGs a natural representation for this information.

Despite such heavy and widespread success of KGs for representing entities in the world and their properties, Taylor (2017) points out that there has not been much attention paid in the research community to *class-level attributes* in KGs: graph edges between nodes that represent categorical terms, what they might mean and how to acquire them. For the purposes of this paper we use the words *type, category, class* interchangeably, as well as *attribute, property, relation*. Practical and industrial KG edges remain almost exclusively at the instance level (e.g., McDonalds serves Big Mac), and a few KGs may encode class-level domain/range constraints (e.g., Restaurants serve Food), but no KG includes attributes of classes that represent our common-sense knowledge about them (e.g., Burger Joints serve burgers). There has certainly been a lot of research published in the sub-fields of Knowledge Representation on axiomatic knowledge acquisition, for example Ji et al. (2020), but these methods are not well-suited for crowdsourcing and have not made the transition to any industrial KG settings.

In this paper we explore the question of acquiring common sense class-level attributes from the crowd and applying those attributes effectively with other sources of information to solve a knowledge-base completion (KBC) problem, as defined in Bordes et al. (2013), where success is measured by the precision and recall of graph edges. We take a particular problem, that of understanding what is *offered* at each *establishment* on earth. Such a KG could be used to answer questions like, "Where can I buy an umbrella nearby?" (see **Figure 1**), "Where can I eat lamyun?", or "Where can I get a flu shot?", etc. We call this problem *local offerings* and it is one that is of interest to search engines like Google.[1]

Local offerings, compared to on-line, poses a significant practical knowledge acquisition problem because real-world transactions do not occur on-line or the data is heavily siloed, and therefore data about what products are being sold at what stores, or what dishes are served at what restaurants, is not broadly available; it is a sort of "dark matter" of the web—we know it's there but can't directly observe it. Although it may seem familiar to us—e.g., brick and mortar shops that support on-line ordering and in-store pickup—such exceptions are actually quite rare, by the numbers. Less than 30% of stores worldwide having a website and even fewer that include a product catalog.[2]

Indeed, our data shows that web pages and merchant feeds account for less than 0.001% of the total matrix of products at stores. To address this shortage of web information, we harness the crowd in three tiers: *users* around the world who have visited stores and voluntarily provide instance-level product availability (e.g., Ajay Mittal Dairy sells Milk); a much smaller set of *paid raters* who curate class-level attributes connecting common sense store and product categories (e.g., Grocery Stores sell Milk); and a very small set of *paid operators* who call stores to confirm the instance-level associations as evaluation ground-truth labels. The intuition behind this combination is that a lot of the instance-level associations are obviously true or false at the categorical level, and that acquiring knowledge at that level can jump-start the instance-level acquisition and help it be more productive: don't waste a user's efforts answering about milk or asphalt at an individual grocery store when simple common sense tells us the answer. Due to the prominence of common sense curation in our approach, we call the project *CrowdSense* (CS).

To our knowledge, acquiring class-level attributes from the crowd in order to jump-start a KBC problem has not been attempted before, and there are very few examples of KBC problems at this scale (tens of millions of stores wordwide and more than 10k products). The project and approach led to a successful worldwide launch of local shopping results overlaid on Google Maps, and involved many complexities beyond the scope of this paper, including more than 2 years of data collection at a worldwide scale. Due to this complexity and scope, we focus here on the real-world knowledge acquisition aspect of the work, and present a few simplified experiments that demonstrate how the acquired knowledge can be used for KBC. The contributions of this paper are primarily:

- To demonstrate that class-level bipartite knowledge acquisition can be effective in approximating instance-level knowledge (Section 5.5) as a solution to *label sparsity*;
- A crowdsourcing approach to acquire such class-level knowledge for the local shopping problem (Section 5.4);
- Experimental results that show the effective combination of class- and instance- level knowledge from various sources used in the launched system (Section 6.3).

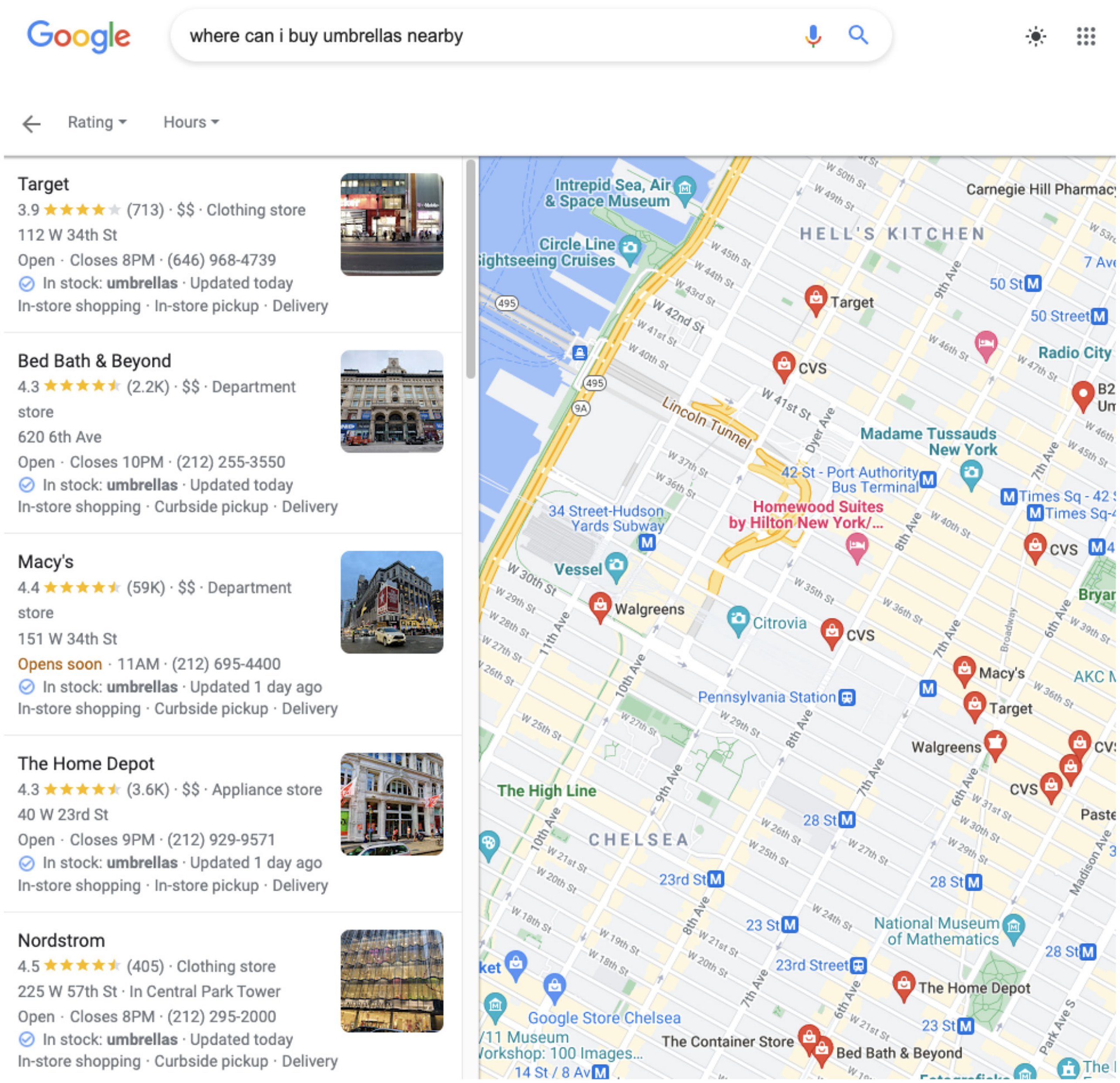
The approach has generalized to other bipartite relations between places and types of entities that are organized in a taxonomy, such as dishes at restaurants, services at professional offices, etc., as well as a wide range of other bipartite graph problems where common sense or categorical knowledge prevails as defaults, such as ingredients for dishes, linnean taxonomies of living creatures, etc.

2. FORMALIZATION

We start with an initial knowledge graph $\mathcal{G}'(\mathcal{I}_P \cup \mathcal{C}, \mathcal{R}_T \cup \mathcal{R}_{SC})$, where $\mathcal{C} = \mathcal{C}_P \cup \mathcal{C}_O$ forms the set of all categories,

[1] https://support.google.com/merchants/answer/9825611?hl=en

[2] https://www.forbes.com/sites/jiawertz/2018/05/17/how-brick-and-mortar-stores-can-compete-with-e-commerce-giants/#2019f5a23cc0



where can i buy umbrellas nearby

Rating ▾ Hours ▾

Target
3.9 ★★★★☆ (713) · $$ · Clothing store
112 W 34th St
Open · Closes 8PM · (646) 968-4739
✓ In stock: **umbrellas** · Updated today
In-store shopping · In-store pickup · Delivery

Bed Bath & Beyond
4.3 ★★★★★ (2.2K) · $$ · Department store
620 6th Ave
Open · Closes 10PM · (212) 255-3550
✓ In stock: **umbrellas** · Updated today
In-store shopping · Curbside pickup · Delivery

Macy's
4.4 ★★★★★ (59K) · $$ · Department store
151 W 34th St
Opens soon · 11AM · (212) 695-4400
✓ In stock: **umbrellas** · Updated 1 day ago
In-store shopping · Curbside pickup · Delivery

The Home Depot
4.3 ★★★★★ (3.6K) · $$ · Appliance store
40 W 23rd St
Open · Closes 9PM · (212) 929-9571
✓ In stock: **umbrellas** · Updated 1 day ago
In-store shopping · In-store pickup · Delivery

Nordstrom
4.5 ★★★★★ (405) · Clothing store
225 W 57th St · In Central Park Tower
Open · Closes 8PM · (212) 295-2000
✓ In stock: **umbrellas** · Updated today
In-store shopping · Curbside pickup · Delivery

FIGURE 1 | Google Maps local shopping search results for umbrellas in NYC shows stores that sell them.

partitioned into place $\{c_p \in \mathcal{C}_P\}$ and offering $\{c_o \in \mathcal{C}_O\}$ categories (e.g., hardware-store, power-tools, resp.), and $\{i_p \in \mathcal{I}_P\}$ the set of all place instances (i.e., the establishments such as stores and restaurants themselves). The edges of the graph are the class/instance (also known as type) relation between place instances and place categories $\{\langle i_p, c_p\rangle \in \mathcal{R}_T\}$, and the subclass relation $\{\langle c_p, c_p'\rangle \in \mathcal{R}_{SC}\}$ with a disjointness constraint

$$\langle x, y\rangle \in \mathcal{R}_{SC} \implies \{x, y\} \subset \mathcal{C}_P \oplus \{x, y\} \subset \mathcal{C}_O$$

so that the relation is only defined over pairs of categories belonging to the same type. Lastly each of these primitive sets are disjoint $\mathcal{I}_P \cap \mathcal{C}_P = \mathcal{I}_P \cap \mathcal{C}_O = \mathcal{C}_P \cap \mathcal{C}_O = \emptyset$, making $\mathcal{G}'$ tripartite. As usual, $\mathcal{R}_{SC}$ forms a partial order within each (place and offering) category partition, and is transitive over the subcategory relation so that $\langle x, y\rangle \in \mathcal{R}_T \wedge \langle y, z\rangle \in \mathcal{R}_{SC} \rightarrow \langle x, z\rangle \in \mathcal{R}_T$. This is meant to capture a traditional kind of knowledge-graph scenario.

Problem 1. *The* local offerings problem *is the extension of $\mathcal{G}'$ to $\mathcal{G}(\mathcal{I}_P \cup \mathcal{C}, \mathcal{R}_T \cup \mathcal{R}_{SC} \cup \mathcal{R}_I \cup \mathcal{R}_C)$ through the addition of the* class-level *offering availability relation* $\{\langle c_p, c_o\rangle \in \mathcal{R}_C\}$ *and the* instance-level *offering availability relation* $\{\langle i_p, c_o\rangle \in \mathcal{R}_I\}$.

The place instances $\{i_p \in \mathcal{I}_P\}$ represent individual physical places like *Trader Joe's at 142 14th St.* (TJ142), each of which

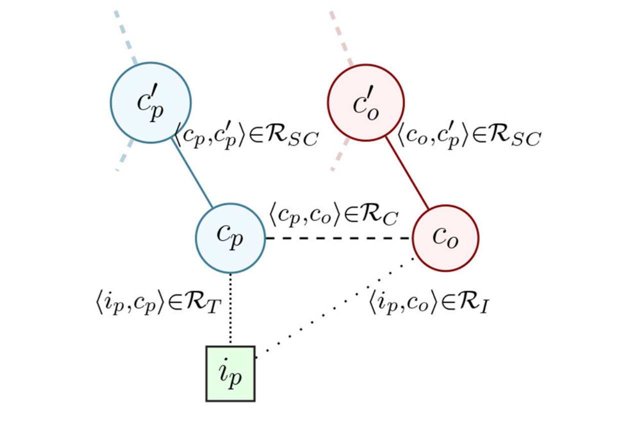

c'_p
$\langle c_p,c'_p\rangle\in\mathcal{R}_{SC}$
c'_o
$\langle c_o,c'_p\rangle\in\mathcal{R}_{SC}$
$\langle c_p,c_o\rangle\in\mathcal{R}_C$
c_p
c_o
$\langle i_p,c_p\rangle\in\mathcal{R}_T$
$\langle i_p,c_o\rangle\in\mathcal{R}_I$
i_p

FIGURE 2 | Example subset of graph $\mathcal{G}$ with a place instance i_p, a place category c_p, its parent category c'_p, a offering category c_o, its parent c'_o and the class- and instance- level offering availability relations between them.

is typed with some number of place categories $\{c_p \in \mathcal{C}_P\}$ like *Supermarket*. The offering categories $\{c_o \in \mathcal{C}_O\}$ represent the types of offerings at all places, such as *Milk* or *Dairy*, so that $\{\langle TJ142, Milk\rangle \in \mathcal{R}_I\}$ means that particular Trader Joe's sells Milk. Note that a more complete definition of the local shopping problem would include the extension of $\mathcal{C}_O$ to instances (i.e., place inventory), but we do not have access to that data, and use this definition as a simplification that serves to answer most *local offering* queries. A simple example is shown in **Figure 2**, showing four categorical graph nodes and one instance node, with each of the relation types shown as edges.

This simplification is best understood as a matrix $\mathbf{R} : \mathcal{I}_P \times \mathcal{C}_O$ representing $\mathcal{R}_I$, where $\mathbf{R}_{i,j}$ are observations (or predictions) that place i offers j. With enough observed $\mathbf{R}_{i,j}$, collaborative filtering methods (e.g., matrix factorization) can be exploited to predict unobserved values from observed ones. Moving between matrix and graph representation can be done in a variety of ways, such as thresholding matrix values into discrete edges in $\mathcal{R}_I$, or using a graph formalism that supports confidence values on edges, as described in Noy et al. (2006).

We argue that the real world grounding of the $\mathcal{R}_I$ association in people's everyday experience allows us to exploit meaningful common sense *categorical* knowledge for the problem of acquiring the edges in $\mathcal{R}_C$, and use simple defeasable methods to then infer the edges in the graph for the relation $\mathcal{R}_I$.

3. VOCABULARY

The local offerings system and all the experiments described in this paper use the open *Google My Business* (GMB) categories[3,4] for place categories ($\mathcal{C}_P$) and *Google Product Taxonomy*[5,6] for the offering categories ($\mathcal{C}_O$). Each set comes with a taxonomic structure that we encode as the $\mathcal{R}_{SC}$ relation, every category has at least one parent category with the exception of the top-level (most general) categories, and a few categories have multiple parents.

This project began with shopping and was extended to dining by adding a number of dishes to $\mathcal{C}_O$. These dishes are from Google's KG, and most of them can be found in Freebase under the type `/food/dish`. The restaurant categories are already part of the GMB set.

There are roughly 15k products categories in $\mathcal{C}_O$, that are similar in semantics to UPCs (Universal Product Code, the bar codes on most packaged products), grounding out in 19 top-level categories. There are roughly 10k dishes in $\mathcal{C}_O$, that are similar to menu items, with very little taxonomic structure. The GMB categories that comprise in $\mathcal{C}_P$ include many that are unrelated to local shopping or dining, so we restrict ($\mathcal{C}_O$) to those below *store* and *restaurant*, resulting in roughly 3k with those two roots.

These taxonomies have different graphical structure: the product taxonomy is fairly deep, and the place taxonomy is fairly shallow, yet they align surprisingly well. For example, there is a deep taxonomy of products under "Grocery," and a store category "Grocery Store." There are a few misalignments, for example "Batteries" are under "Electronics" but are sold at "Drugstores." A few of these misalignments are ameliorated by hybrid categories like "Household products," which is an additional ancestor for "Batteries." The food taxonomy we used from Freebase is nearly flat, making for an interesting comparison on the usefulness of a good taxonomy. Note that we do not change the taxonomies or memberships; as defined in Section 2, we treat the initial graph $\mathcal{G}'$ as given.

Finally, Google Maps has tens of millions of establishments worldwide that form the set of places $\mathcal{I}_P$; each has a category label which is displayed in the maps UI under the place name and user rating, giving us the edges in $\mathcal{R}_T$. A large part of these labels are assigned by merchants, some by users, some by operators and others by machine automation. These labels are generally high quality, with precision over 0.8. The largest source of inaccuracies are store labels that are more general than they need to be, when a more appropriate category exists. The labeling infrastructure requires a single "primary" category, while many places could be categorized in several ways. A Glossary of terms defined in this paper has been provided in **Table 1**.

4. A THREE-TIERED CROWD

The system for which we performed the crowdsourcing described in this paper is quite large and complex, and is launched and available to users worldwide through search. It uses a DNN model to predict $\mathcal{R}_I$ pairs from many signals that include information extraction (IE) from store web pages, direct merchant feeds, store type, and dozens of other features that include a significant amount of user-generated content (UGC).

The well-known bipartite problems that have been solved by machine learning have the advantage that the organizations that solved them had a lot of labeled data for those problems. For example, Netflix has millions of ⟨user, movie⟩ pairs, and can use this massive data to seed big machine learning systems to better predict what movies a user make like. A vast number of practical bipartite problems, however, have very little data, resulting in *label sparsity*.

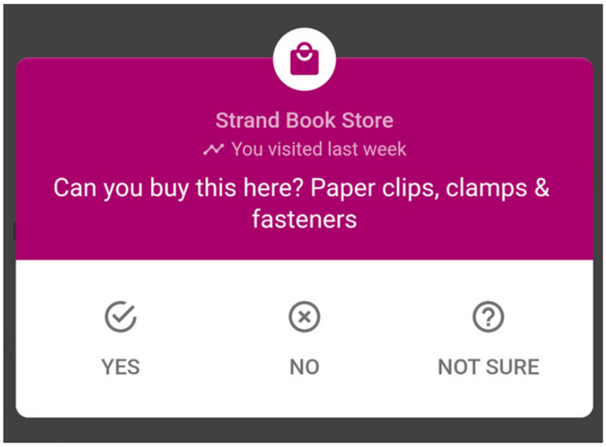
[3] https://support.google.com/business/answer/3038177/#categories
[4] https://bayareawebsitedesigner.com/gmb-categories/
[5] https://www.google.com/basepages/producttype/taxonomy.en-US.txt
[6] https://feedonomics.com/google_shopping_categories.html

TABLE 1 | Glossary of terms.

Terms	
Place	An establishment (store or restaurant) on Google Maps
Offering	A product or dish available at a place
KBC	Knowledge Base Completion
GMB	Google my Business (source store categories)
GPT	Google Product Taxonomy (source product categories)
UGC	User Generated Content–user responses to yes/no questions
CS	Crowd Sense, our approach
WebIE	Information extraction of offering names from place web pages
WALS	Matrix factorization using WALS to predict $\langle i_p, c_o \rangle$ pairs
Knowledge graph	
$\{i_p \in \mathcal{I}_P\}$	Set of place instances
$\{c_p \in \mathcal{C}_P\}$	Set of place categories
$\{c_o \in \mathcal{C}_O\}$	Set of offering categories
$\langle c_p, c'_p \rangle \in \mathcal{R}_{SC}$	Place subclass/superclass relation
$\langle c_o, c'_o \rangle \in \mathcal{R}_{SC}$	Offering subclass/superclass relation
$\langle i_p, c_p \rangle \in \mathcal{R}_T$	Place instance/class type relation
$\langle c_p, c_o \rangle \in \mathcal{R}_C$	Class-level offering @ place availability relation
$\langle i_p, c_o \rangle \in \mathcal{R}_I$	Instance-level offering @ place availability relation
$\mathcal{G}'$	Base KG of place/offering classes and place instances
$\mathcal{G}$	$\mathcal{G}'$ extended with $\mathcal{R}_C$ and $\mathcal{R}_I$
$\mathbf{R}_{i,j}$	Likelihood that place instance *i* sells offering class *j*
Crowd task	
$w_{x,o}$	Rater score for place (class or instance) x and offering class o
$\alpha_{c,o}$	Number of "always" answers for class-level pair $\langle c, o \rangle$
$\nu_{c,o}$	Number of "never answers for class-level pair $\langle c, o \rangle$
$y_{i,o}$	Number of "yes" answers for instance-level pair $\langle i, o \rangle$
$n_{i,o}$	Number of "no" answers for instance-level pair $\langle i, o \rangle$

Label sparsity means that machine learning systems don't have enough data to make reasonable predictions, and the only way to move forward is to acquire it. Acquiring the data needed to seed large scale AI systems is as much a problem today as it was during the bygone era of expert systems, where, according to Shortliffe and Buchanan (1975) and many others, the bulk of the research focus was on algorithmic solutions to rule-based reasoning problems, but the bulk of the difficulty and work was in knoweldge acquisition. This history continues to repeat itself; Sambasivan et al. (2021) point out that knowledge acquisition is viewed as less glamorous than inventing new neural algorithms and architectures. As noted above, for the local shopping and dining problems, existing sources gave us less than 0.001% of the total matrix **R**, leaving a huge knowledge acquisition problem. We developed a novel three-tiered crowd to gather the data discussed in this paper:

- **CrowdSense** (CS): We collected 25k class-level $\langle c_p, c_o \rangle \in \mathcal{R}_C$ pairs for shopping and 20k for dining, from a pool of paid raters. Though a relatively small crowd effort, this ends up being the largest source of instance-level $\langle i_p, c_o \rangle \in \mathcal{R}_I$ pairs through default inference (full details in Section 5), yielding billions of instance-level pairs.
- **UGC**: Google Maps provides the facility for users to voluntarily add reviews, photos, venue categorization, and attributes (e.g., "has Wi-Fi") to places they've visited. Through the UGC framework, users answer yes/no questions about product and dish availability at places they've visited, shown in **Figure 3**. While Google's deployed local search system does use all the UGC data, including reviews and photos, etc., in this paper we only describe and analyze the impact of the yes/no questions, which comprise the largest crowdsourcing element of the system, at millions of answers per day. Each user is given a set of $\langle i_p, c_o \rangle$ pairs to answer, giving us a distribution of yes and no answers for each pair. In the experiments shown in Section 6, we show the growth in coverage over time as more answers are collected, yielding hundreds of millions of instance-level pairs over the course of this study (2 years for shopping and 15 months for dining).[7]
- **Gold**: We collected 40k gold standard $\langle i_p, c_o \rangle$ pairs for shopping, and 20k for dining, by having paid operators call each place i_p and ask them if they sold or served c_o. The places were selected from among more than 50 countries with the top-5 countries being US (20%), JP (5%), IN (5%), GB (5%), BR (4%); places within each country were sampled uniformly to provide a microcosm of representative demographics. Clearly the highest fidelity and most expensive data, it is by far the smallest.



Strand Book Store

You visited last week

Can you buy this here? Paper clips, clamps & fasteners

FIGURE 3 | Example question used to gather UGC.

One of the critical obstacles to gathering this data from people in all tiers is the class imbalance: less than 4% of the possible store-offering pairs are positive. Gathering 96% negative results is a waste of human labeling resources and, far more critical, makes for an unsatisfactory user experience—users want to feel helpful and answering 9/10 negative questions is frustrating. Moreover, particularly obvious negative questions, like fish heads at a hardware store, confuse some users into saying they are unsure—the questions are so obvious they feel they must be missing something. Finally, a few of these obvious negatives end up on social media as jokes, which is embarassing.

Active learning (AL) is a known method for dealing with class imbalance—sampling near the classifier boundary typically

[7] Collection continues, these windows were used for this paper.

yields a good balance between positives and negatives and, thus, provides utility for training the model. Unfortunately for problems where there is also label sparsity, there is not enough data to train a model and so nothing to base AL on. Our class-level approach offers a solution to this problem as well. As described in more detail in Section 5, we gather a distribution of judgements on class-level pairs, and the resuling pairs fall into three categories: *obviously available* (e.g., grocery store, milk), *obviously unavailable* (e.g., hardware store, fish heads), and *possible* (e.g., hardware store, 9 inch nails). The *possible* category of class-level pairs captures products that are available at some, but not necessarily all, stores in the class, and provide excellent guidance for selecting instance-level pairs to ask users.

Even after we'd acquired enough data to begin training a model and use AL, the *possible* category offered an additional benefit. In the early stages of acquiring training data, known as the explore (vs. exploit) stage, $\langle i_p, c_o \rangle$ pairs with enough evidence to be close to the classifier boundary are very likely to be positive, so much so that the class balance of margin sampling was 80% positive. Clearly a 50% class balance could then be achieved by up-sampling pairs that are further below the classifier boundary, however such an approach is very likely to choose these problematic obvious negatives discussed above. A mix of possibles with margin sampling was able to achieve a 50% class balance with high utility and no embarassment.

For the Gold data, class imbalance presents as much of a problem as for UGC, however since this data set is used to measure the quality of the CS data, we did not want to bias our evaluations by using CS as a guide. Instead, to achieve better class balance, the WebIE baseline data (q.v. below) was used to guide the collection toward pairs that had an increased chance of being true; for example, if a places's webpage mentioned an offering we would try to call places of the same type and ask about that offering. We enforced a positive/negative class balance of 50%, and targeted a stratification of the sampling that preserved the 30/70 balance of places with and without websites.

5. CROWD SENSE

The obvious way to gather the edges in $\mathcal{R}_I$ would be to use store inventory or transaction records. The problem with this approach is that *local offerings* is still mostly an off-line or highly siloed process worldwide, and we did not have access to transactional data that gives us these observations. Google provides merchants a free way to share their menus or inventory on-line, but much fewer than 1% of places worldwide had made use of it. Our data showed that web pages and merchant feeds together accounted for less than 0.001% of the space of the matrix **R**, giving us the label sparsity problem. Filling the cells of matrix **R** means acquiring the edges in $\mathcal{R}_I$, and we propose to accomplish this by starting with the acquisition of edges in $\mathcal{R}_C$, the *class level attributes*, and inferring those values as defaults for $\mathcal{R}_I$.

5.1. Crowd Hypothesis

The intuition driving our approach is that the crowd can provide the class-level knowledge ($\mathcal{R}_C$) by appealing to their common sense experience; everybody knows that, e.g., "All supermarkets sell milk." Reality is more complicated, and since the problem space is sparse, the class-level data is also dominated by what offerings are obviously *not* available. Far behind the obvious negatives are, as discussed above, the possibles—offerings that are *usually*, but not always, available at some type of establishment. Wasabi Peas, while they are found almost exclusively in grocery stores, are not found in all of them. What we really aim for the crowd to provide is a *distribution* of the offerings available at places of a given type. This is where a lot of existing knowledge graph methods fail, especially at the class-level, as they rely on an assumption of discreteness.

It may seem that we could ask individual people to answer a question like, "What percent of stores of type c_p sell product c_o?" However, research in human computation such as Surowiecki (2005) has shown that individuals cannot reliably answer such questions. Using (Welty et al., 2012; Aroyo and Welty, 2014, 2015) as a starting point, we hypothesized:

Hypothesis 1. *Asking multiple raters about the same categorical pairs would produce a distribution of answers that approximate the real world distribution of* $\mathcal{R}_I$.

In other words, if 70% of raters say that oat milk is sold at grocery stores, then 70% of grocery stores will sell oat milk.

Before testing our hypothesis, we ran numerous pilots to tune the hyper-parameters of the crowd task in the shopping domain, asking raters questions about 11k $\langle c_p, c_o \rangle$ pairs from 154 store types and 3600 products in five countries. We experimented with: the number of raters per pair, testing between 5 and 25 raters per pair; the size of the rater pool, ranging from 100 to 500; the question phrasing; and the answer options. Based on manual analysis of the cost and quality, we settled on these task hyper-parameters: five raters per pair, randomly selected from a pool of 130 raters in six countries, sourced from contracted operators through an in-house crowdsourcing platform, and the question, "Would you expect to find c_o products in stores of the category c_p?" with four answer options ("Always Available," "Sometimes Available," "Never Available," "I don't know"). For dishes, the question was rephrased, "Would you expect to find c_o dishes in restaurants in the category c_p?"

Under these settings, our final PRODCAT task (see below) gathered 25k class-level ($\langle c_p, c_o \rangle$) pairs with 5 labels per country, that through inference (*q.v.* Section 6.1) resulted in billions of $\langle i_p, c_o \rangle$ pairs, 99% of which were negative. It took 6 weeks to run and analyze the pilots, and 2 weeks to run the final task. For dishes, the MATRIX task collected 15k class level pairs from 5 raters per pair in 2 weeks, resulting in billions of instance-level pairs.

Raters were supplied by a set of contractors who are obligated to follow Google's Code of Conduct, and were managed by an administrator outside our group. The MATRIX and PRODCAT task designs (q.v. below) grouped between 200 and 400 pairs in a single matrix, raters were assigned a matrix by the administrator based primarily on availability. Many raters were assigned multiple matrices over time, but in our analysis we did not account for individual characteristics of raters (such as expertise), even though we know from Aroyo and Welty (2014) this can yield improvements.

5.2. Data Collection Tasks

Another way to state our hypothesis is that the categorical crowd disagreement should reflect the real world distribution, but disagreement can have many causes that are not related to the desired distribution. The various pilot tasks we ran represented a gradual refinement of the data and task descriptions to eliminate disagreement from other causes. We report here on four different approaches for the shopping domain:

5.2.1. RANDOM

To confirm the sparsity of $\mathcal{R}_C$, we randomly and independently selected category pairs from $\mathcal{C}_P \times \mathcal{C}_O$, weighing the selection from $\mathcal{C}_P$ proportionally to the number of stores belonging to each category (i.e., larger categories are more likely to be selected). Pairs were presented to 5 raters from the same country. This RANDOM task confirmed that the vast majority of pairs are "obvious" negatives (asphalt at grocery stores, cars at violin shops, etc.), as more than 95% of the pairs resulted in 5 "Never" ratings.

5.2.2. SINGLETON

To address the sparsity shown in RANDOM, we leveraged web signals (see Section 6.1) to select pairs with more likelihood to be available at places within a given category, and presented one pair at a time to 5 raters from the same country. This resulted in a distribution of rating scores ranging from all-5 "Always Available" to all-5 "Never Available" skewing toward the positive (always) side. The SINGLETON task results showed disagreement from other causes, described in Section 5.3.

5.2.3. MATRIX

To address the disagreement due to ambiguity (Section 5.3), we designed a novel matrix presentation of class-level pairs, with four $\{c_o \in \mathcal{C}_O\}$ as the columns and a set of 100–200 $\{c_p \in \mathcal{C}_P\}$ as the rows, depending on our ability to match offerings to the place categories using web signals. **Figure 4** shows the matrix presentation (with data sampled through the PRODCAT method below). The advantage of this presentation is that raters familiarized themselves with a category and answered many questions related to it, rather than having to understand one pair at a time. This approach still produced some unwanted disagreements due to difficulty understanding some of the products, esp. very specific ones, and we were concerned that the web signals were biasing our sample toward availability patterns of online places, rather than our target class of establishments without web pages. Most importantly, the amount of time the raters spent per $\langle c_p, c_o\rangle$ dropped by 50%.

5.2.4. PRODCAT

The final crowdsourcing task used the MATRIX presentation but changed to a dynamic method that sampled the $\langle c_p, c_o\rangle$ pairs starting at the top of the product taxonomy, and working down the $\mathcal{R}_{SC}$ relation from most general to most specific. It was not useful to treat the store taxonomy this way, as it is very shallow, and we did not have a dish taxonomy. When a pair was given an overall negative label, we did not sample any subcategories of c_o and inferred a negative label for all descendents. For example, since *Auto parts stores* do not sell *Grocery* and $\langle Dairy, Grocery\rangle \in \mathcal{R}_{SC}$, we did not ask ⟨*Auto parts stores, Dairy*⟩.

The product taxonomy is not a strict tree, but a DAG, and when reconciling conflicting ratings from multiple parents, we retained the most positive rating. *Electronics* are not sold at *Pharmacies*, whereas *HouseholdProducts* are sometimes sold there, and *Batteries* are a subcategory of both *Electronics* and *HouseholdProducts*, so we do ask about ⟨*Batteries*, *Pharmacies*⟩.

This top-down taxonomic pruning eliminated any need for the web signals, and accounted for the sparsity at a very high level, since (by accident or ontology) the store and product categories were well aligned: e.g., *Auto parts stores* sell *Auto parts* and do not sell *Groceries*. Higher level categories also made a lot more sense to raters when presented with a sub-category, e.g., *Sports and Outdoor Electronics* with *Fitness Trackers*, and since our rater pool did not vary much, they became familiar with the taxonomic distinctions as they progressed down the taxonomy, which was evidenced by a reduction in visits to the taxonomy element descriptions over time.

5.2.5. Dish MATRIX

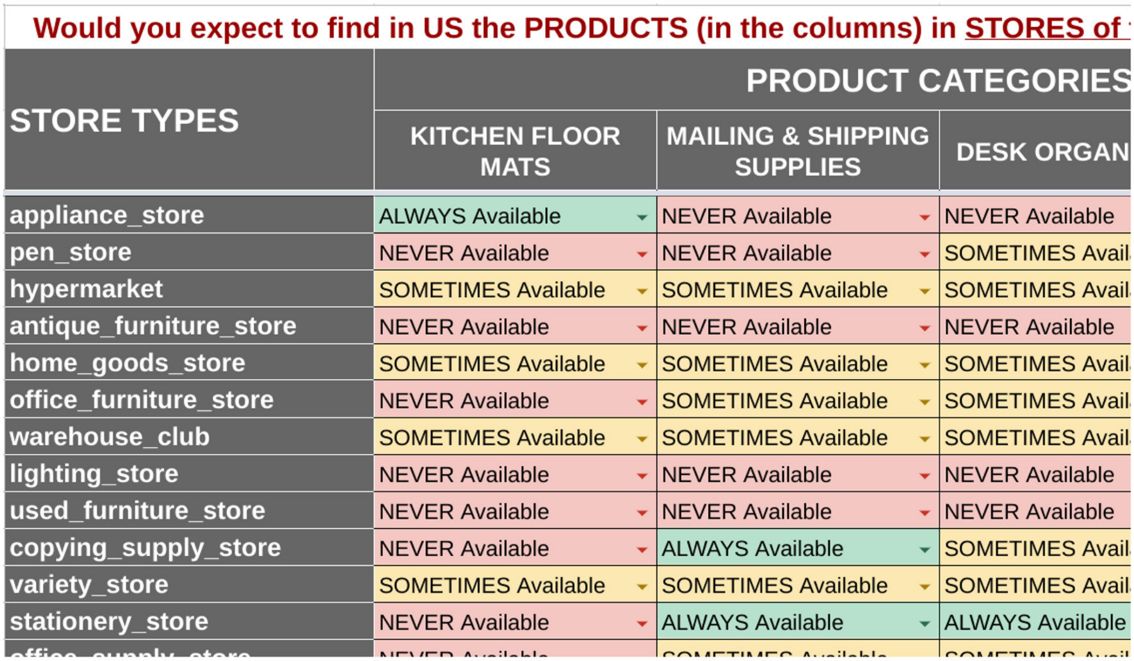
To gather the class-level pairs $\langle c_p, c_o\rangle$ for the dining domain, we were not able to fully reuse the PRODCAT method, since the dishes in our KG did not have taxonomic organization, which was the key to the improvements of PRODCAT over MATRIX. Instead, we used the MATRIX method, presenting the class-level pairs in a matrix, selected by their popularity in web signals. As with singleton, this approach favored positive pairs, indeed our raters appear to have been overly positive in their answers.

5.3. Ambiguity

In the pilot experiments run for shopping we observed disagreement in the results that did not support our crowd hypothesis, but were caused by ambiguity such as:

- product is a material, substance (e.g., plastic, starch, arugula) or some product aspect (e.g., color, size)
- product is a brand (e.g., Avian, Kleenex) or contains a brand name (e.g., Nike Sneakers, Todd's boots)
- place or offering is too specific (e.g., duck sauce, goat meat, vanilla orchids, banner store)
- place or offering is too generic (e.g., gift, organic food, chicken, restaurant)
- offering is regional (e.g., Harissa, Jajangmyeon)
- offering is seasonal (e.g., christmas trees, flip-flops)
- offering is polysemous in a way that is resolved by the store type, e.g., "fish" in a grocery store vs. a pet store
- flashy menu item (e.g., nacho fries bellgrande, del monde delux).

In MATRIX and SINGLETON, for example, raters seem more willing and able to answer the question, "Is milk sold here?" compared to "Is dairy sold here?" In the latter case, there is uncertainty over what minimum set of dairy items (milk, cheese, butter, yogurt, etc.) would be needed for "sells dairy" to be true, yet the equally rich sub-categories of milk (whole milk, skim milk, organic milk, etc.) did not cause the same uncertainty. When presented with the categories in a top-down fashion, raters



Would you expect to find in US the PRODUCTS (in the columns) in STORES of

STORE TYPES	PRODUCT CATEGORIES		
	KITCHEN FLOOR MATS	MAILING & SHIPPING SUPPLIES	DESK ORGAN
appliance_store	ALWAYS Available	NEVER Available	NEVER Available
pen_store	NEVER Available	NEVER Available	SOMETIMES Avail
hypermarket	SOMETIMES Available	SOMETIMES Available	SOMETIMES Avail
antique_furniture_store	NEVER Available	NEVER Available	NEVER Available
home_goods_store	SOMETIMES Available	SOMETIMES Available	SOMETIMES Avail
office_furniture_store	NEVER Available	SOMETIMES Available	SOMETIMES Avail
warehouse_club	SOMETIMES Available	SOMETIMES Available	SOMETIMES Avail
lighting_store	NEVER Available	NEVER Available	NEVER Available
used_furniture_store	NEVER Available	NEVER Available	NEVER Available
copying_supply_store	NEVER Available	ALWAYS Available	SOMETIMES Avail
variety_store	SOMETIMES Available	SOMETIMES Available	SOMETIMES Avail
stationery_store	NEVER Available	ALWAYS Available	ALWAYS Available
office_supply_store	NEVER Available	SOMETIMES Available	SOMETIMES Avail

FIGURE 4 | Partial view of the PRODCAT data collection template with example answers from one rater.

first dealt with their uncertainty about "dairy" and applied it to the subcategories as well, and this handled most of the general and specific ambiguity, and for many store types, raters were willing to give definite answers about the other sub-types in subsequent tasks.

We specifically addressed the material, aspect and brand problems by removing them from the product set, their treatment is the subject of future work. We instructed the raters to treat seasonal products as "year round," after confirming that users are less likely to search for such products out of season. We updated the task design to allow raters to explore the two taxonomies to help with polysemy, but we found that grouping store categories by taxonomic (sibling and parent) relations in PRODCAT obviated this exploration.

Regional products produced disagreement esp. across countries, where for the final tasks we sourced raters in six countries (US, IN, BR, FR, JP, IN). Often this showed up merely as "I Don't Know" answers which were not used in predicting $\mathcal{R}_I$, but do show up in IRR. More interesting cases included when a product had a slightly different meaning, or was sold in different types of stores, in different regions. For example, "syrup" in France is sold in drug stores, and raters in other countries did not agree. This is because in France "syrup" is cough syrup, and this association did not exist elsewhere that we tested. We had many expectations for the role of, and differences between, raters in different countries, described in more detail in Section 5.6. Despite these anectdotal examples, class-level ratings from one country were generally worse at predicting instance-level availability within the same country, and better at predicting other countries. In the final system, we ignored the country of the class-level ratings, treating all raters as equal.

Flashy menu items, in which superlatives and other postive-sentiment modifiers are added to dish names, were an additional problem in dining that we did not observe in the shopping domain. This is in part due to the taxonomy curation of the shopping data, in which such modifiers had been removed to create a fairly neutral set of categories. For dining, which lacked the taxonomy, some raters were able to identify the superlatives as meaningless, or were familiar with the dish names because they came from well-known chains, while other raters didn't have that knowledge and would answer either negatively or uncertainly. Our scoring method effectively neutralizes such dish names (see Section 5.5), as the disagreement moves the score close to zero, and we did not choose to address it otherwise. Our current work seeks to address this problem through the automatic development of a taxonomy.

5.4. PRODCAT Data Collection Task

The final design of the PRODCAT task, which was only used in the shopping domain, presented a matrix of $\langle c_p, c_o \rangle$ pairs to raters in six countries, five raters per country, and consisted of several elements:

- a list of store categories, $c_p \in \mathcal{C}_P$
- a list of product categories, $c_o \in \mathcal{C}_O$
- c_p, c_o pairs presented in an $n \times 4$ matrix, where each c_p is a row and each c_o is a column; n ranged from 40 to 200 depending on our ability to find suitable products

TABLE 2 | Example CrowdSense ratings on $\mathcal{R}_C$ pairs.

Category	Product	Always	Some	Never
Auto parts store	Pita	0	0	5
Bakery	Longline Vests	0	0	5
Beauty supply store	Aromatherapy	5	0	0
Bicycle store	Home furnishings	0	0	5
Butcher shop	Quicklime	0	0	5
Chinaware store	Watches	0	0	5
Clothing store	Women's shirts	5	0	0
Clothing store	Petite negligee	5	0	0
Clothing store	Truck tailgate caps	0	0	5
Clothing store	Chameleon	0	0	5
Clothing store	Typewriter ribbon	0	0	5
Coffee store	Instant coffee	4	0	1
Cosmetics store	Non-dairy milk	0	0	5
Drugstore	tarragon	0	0	5
Electronics store	Canister vacuums	5	0	0
Feed store	cybex	0	0	5
Fresh food market	Work dresses	0	0	5
Fruits and vegetables	Turkey sausage	0	1	4
Furniture store	Canopy beds	4	1	0
Furniture store	Box springs	4	0	1
Grocery store	Smart light bulbs	0	0	5
Grocery store	Frozen clams	5	0	0
Grocery store	Soy nuts	4	1	0
Home goods store	Storage baskets	4	1	0

- the matrix was prefaced with: "Would you expect to find in *country* the products (in the columns) in stores of the types (in the rows)?"
- each cell in the matrix connected one pair with four possible answers: "Always available," "Sometimes available," "Never available," and "I Don't Know"
- the row and column headers c_o and c_p included links to an image, a short description, and the position in the respective taxonomy
- raters were encouraged to explore the taxonomies in order to better understand categories
- The column product types were chosen such that three were taxonomy-related (sibling or more-specific child) and one was not, e.g., "aspirin," "notebooks," "paper supplies," and "lined paper."

The final matrix PRODCAT crowd template is shown in **Figure 4** with an example of answers provided by one rater. Based on rater feedback and metrics shown in Section 5.5, this presentation helped resolve many forms of polysemy mentioned in Section 5.3.

5.5. Error of Class-Level Ratings

Table 2 shows a small sample of the CS task results for $\mathcal{R}_C$ pairs; we have intentionally downsampled the "5-never" pairs to show a mixture of different vote ratios.

In Welty et al. (2021) we showed that inter-rater reliability (IRR) cannot reflect the quality of ratings where disagreement is the desired result, so we report the *error* of different $\mathcal{R}_C$ pairs in predicting the distribution of $\mathcal{R}_I$ pairs, by comparing ratings-based scores on $\mathcal{R}_C$ pairs against UGC scores on $\mathcal{R}_I$ pairs obtained from users (see Section 4). Each class and instance level pair has a score:

$$w_{x,o} = \begin{cases} (\alpha_{x,o} + \frac{1}{2}\sigma_{x,o})/(\alpha_{x,o} + \nu_{x,o} + \sigma_{x,o}) & \text{if } x \in \mathcal{C}_P \\ y_{x,o}/(y_{x,o} + n_{x,o}) & \text{if } x \in \mathcal{I}_P \end{cases}$$

where $\alpha_{x,o}$ is the number of "always" answers for class-level pairs $\langle x, o\rangle$, $\sigma_{x,o}$ the number of "sometimes," and $\nu_{x,o}$ the number of "never" answers; and $y_{x,o}$ is the number of "yes" answers for store instance-level pairs $\langle x, o\rangle$ and $n_{x,o}$ the number of "no" answers.

Next let $\mathcal{I}_c = \{i : \langle i, c\rangle \in \mathcal{R}_T\}$ be the instances of place category c under R_T. The mean absolute error of class-level pair $\langle c, o\rangle$ is:

$$\textsc{mae}(\langle c, o\rangle \in \mathcal{R}_C) = \frac{\sum_{i \in \mathcal{I}_c} |w_{i,o} - w_{c,o}|}{|\mathcal{I}_c|}$$

The idea is that if the class-level scores ($w_{c,o}$) are an accurate prediction of the availability distribution at the instance level, then they should model user observations at individual stores ($w_{i,o}$), averaged over the size of the store category ($|\mathcal{I}_c|$). **Figure 5** shows the distribution of MAE scores per category pairs for the three shopping and one dining data collection tasks. Despite PRODCAT being a harder task for raters due to the sampled pairs, it performs much better than the other shopping tasks, with nearly half of its categories scoring in the lowest error range, clearly supporting our crowd hypothesis: the disagreement on $\langle c_p, c_o\rangle$ pairs approximates the distribution of $\langle i_p, c_o\rangle$ when $\langle i_p, c_p\rangle \in \mathcal{R}_T$, according to user observations. For Dining, we only ran the MATRIX task, to replicate as much as possible the results from Shopping. As expected, the MAE is lower than for PRODCAT on shopping, but considerably better than MATRIX for shopping. One explanation for this is that our raters were more familiar with dining around the world than shopping, and there was less disagreement caused by not understanding the pair.

5.6. Error of International Ratings

Another hypothesis we formed early on was that raters in our class-level rating pool, which was international, would know their own countries better than other countries, and the initial design of the system called for increasing the weight of in-country class-level ratings over out-of-country ratings when calculating $w_{x,o}$ (see above). In our analysis of CrowdSense errors in the pilot studies, we certainly saw examples of raters misunderstanding dishes and products from other countries (see Section 5.3).

This hypothesis was mostly supported by our analysis of the shopping data, but it turned out to be largely false for dining, to our great surprise, as shown in **Figure 6**; as with **Figure 5**, the charts show the distribution of the normalized MAE from CrowdSense predictions, but in each chart we've restricted the actual restaurants to those within the indicated country, and calculated the $w_{x,o}$ scores for CrowdSense for raters in the country (solid blue bars) and for raters not in the country (hashed red bars). With the exception of Japan, outside raters have a lower error rate, as their distributions are shifted significantly to the left.

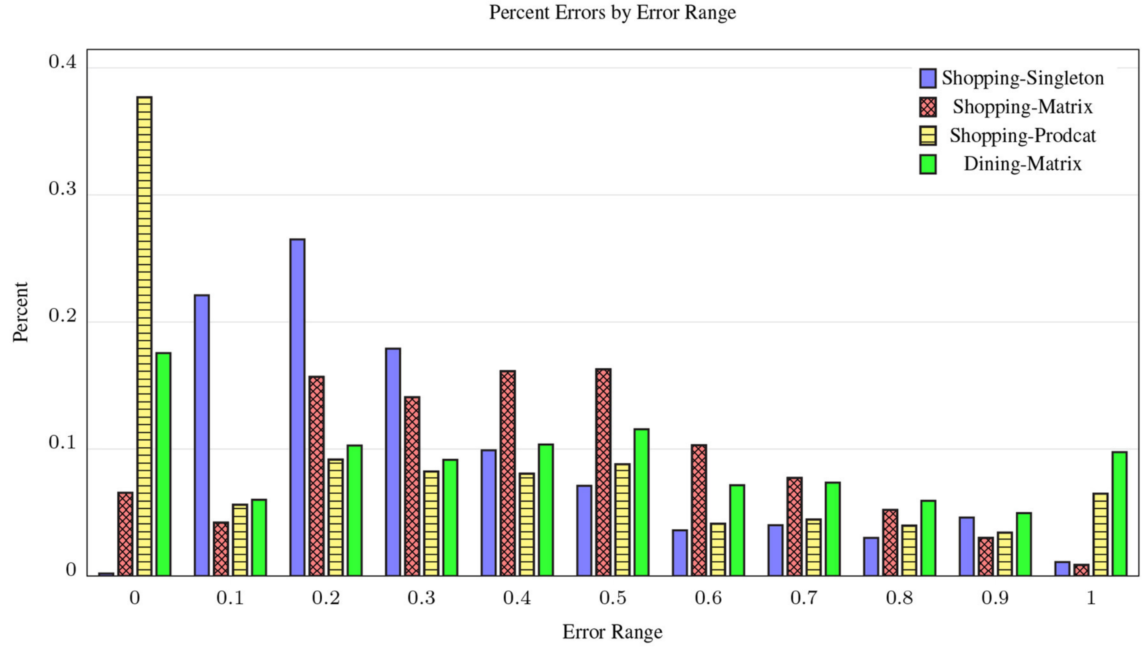


FIGURE 5 | Histogram of Normalized-MAE on CrowdSense pairs for three shopping and one dining (Section 5.5) class-level crowd task designs. Bins to the left indicate the relative number of pairs with lower error, making Shopping-PRODCAT the clear leader. Dining-MATRIX performs better than shopping MATRIX.

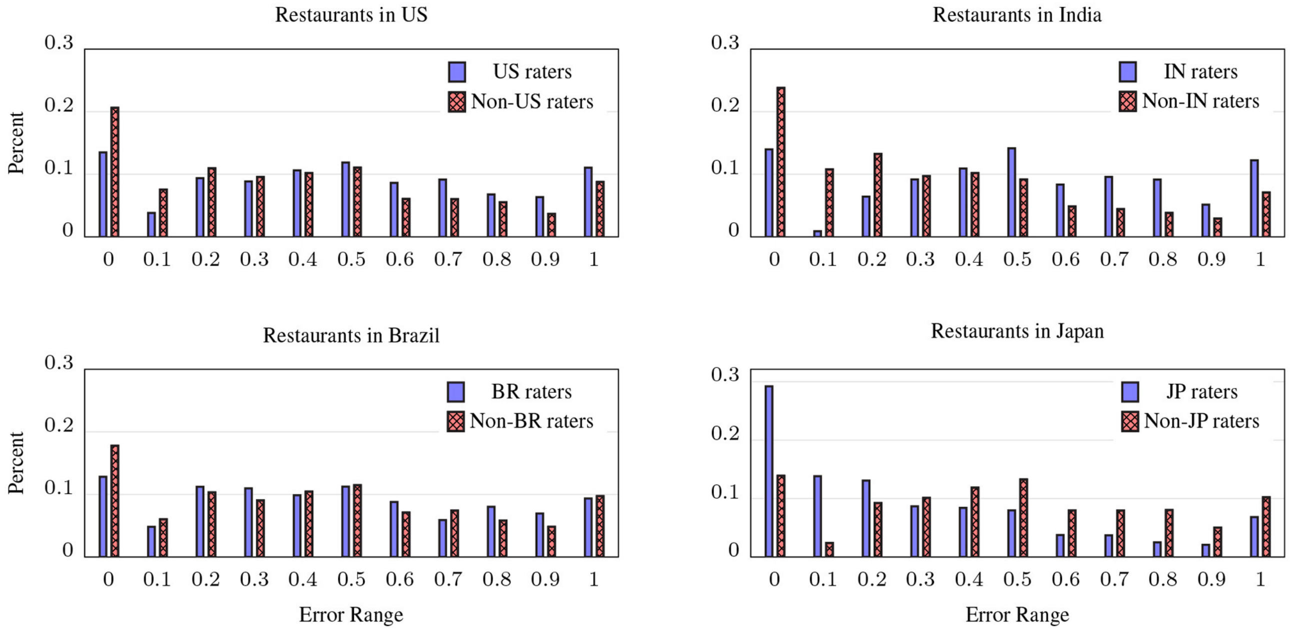


FIGURE 6 | Distribution of CrowdSense errors (Normalized-MAE) for ratings in four countries, comparing CrowdSense predictions from raters in each country to raters outside that country. A shift of scores to the left indicates lower overall error; surprisingly, for all countries except Japan, out-of-country CrowdSense raters are more accurate than those within the country.

In Brazil, the effect is small, in the US it is large and in India the largest. In Japan, the expected effect is dramatic—Japanese CrowdSense raters were far better at predicting the distribution of dishes at Japanese restaurants than non-Japanese raters. We ran the experiment for Germany and Indonesia (not shown) with similar results as the US and India.

For the US, this may be explained by the fact that there are far more chain restaurants that dominate the numbers when calculating the MAE, and many of these chains are familiar abroad, so while US raters are making their decisions based on a broader perspective of chains and non-chains, non-US raters are making their decisions based only on chains, and these capture a larger piece of the US restaurant landscape. In addition, the US has far more restaurants serving international cuisines than any other country, making it possible for international raters to know something about more US restaurants. For Japan, more than any other country, there are many restaurants that serve only a very specific kind of food, and this is well known in Japan and not as much outside it. A possible explanation for the counter-intuitive results in the other countries is that the restaurant taxonomy does not cover those regions very well, leaving more restaurants mis-categorized.

6. INSTANCE-LEVEL PREDICTION EXPERIMENTS

6.1. Data Sources

We compare and contrast several approaches for acquiring and predicting the relations in $\mathcal{R}_I$:

CrowdSense (CS): Class-level associations $\langle c_p, c_o \rangle \in \mathcal{R}_C$ and an associated score for each pair w_{c_p,c_o}, collected through PRODCAT (as described above) for shopping, and MATRIX for dining. In our experiments, we treated the CS data as a static set, although in practice it could grow or change over time like UGC. We collected 25k pairs in the shopping domain and 20k for dining.

User Responses (UGC): As described in Section 4, we collected more than 100M instance-level pairs for shopping from volunteer users around the world over a 2 year period, and roughly half that amount over a 15-month period for dining. Most of the UGC pairs have a distribution of yes and no answers, and more sophisticated processing of the answers is possible, but for simplicity we use the majority vote as the label in the experiments below, where we break the data into sets representing the first $n \in [1, 24]$ months of collection, to illustrate the growth of the data over time.

Web baseline (WebIE): The baseline approach to supporting *local* queries is the Web: using product or dish names mentioned on each place's registered web site as part of an inverted index that are matched to search queries for those products. As discussed above, this approach for local shopping is limited by the coverage of local (aka brick and mortar) stores and restaurants on the web, which was under 30% (60% for the US) at the start of this project in 2017, and has not increased substantially in the years hence. We used a named entity recognizer to extract instance-level pairs ($\mathcal{R}_I : \mathcal{I}_P \times \mathcal{C}_O$) for places with a web site that mention offerings on any of the site's pages, and used the extraction confidence probability threshold yielding 80% precision. WebIE is only able to obtain positive labels, leaving negatives to be inferred from the complement. We chose the 80% precision threshold as this is roughly the precision of the CS inferred data (see **Figures 7**, **8**), which we compare to this and other data sources. While other Web sources (user reviews, coupons, photos, search keyword click-throughs, etc.) and more advanced entity extraction techniques such as Wang et al. (2020) might improve the recall, for most places this information simply is not available. We treated the Web as a single unchanging dataset; for our experiments, the change over time was not significant enough to measure.

WALS(UGC): Since predictions of the instance-level pairs form a matrix, **R**, an obvious approach is to use matrix factorization on the matrix formed from data gathered using the above methods. We used an off-the-shelf WALS implementation based on Koren et al. (2009) trained on the UGC scores discussed below. Since WALS does not use "features," but rather a matrix of real values, we did not include other inputs to WALS in **Figure 7** or **Figure 8**.

6.2. Evaluation

Ultimately our goal is to enable offering queries like, "where can i buy a raincoat?" or "where can i get sesame chicken?" to return nearby places on maps as well as (web) search results; however, direct application impact metrics from our system, which launched in mid-2020, are proprietary. Here we focus on the knowledge acquisition part of the system using metrics of knowledge-based completion, see for example (McNamee and Dang, 2009; Welty et al., 2012).

We collected 40k gold standard $\langle i_p, c_o \rangle$ pairs for shopping, and 20k for dining, by having paid operators call each place i_p and ask them if they sold or served c_o (see Section 4). We used these pairs as a test set in the experiments below. When evaluating against the gold standard, any instance-level pairs that are present in the gold set but missing in the evaluated data are counted as false negatives toward recall. **Table 3** shows a small sample of the shopping gold standard pairs, and **Figures 7**, **8** show the results on 24 and 15 months of UGC data, resp. Note that since WebIE was used to guide the collection of the gold standard, it has a slight advantage in the evaluation.

6.3. Results

6.3.1. WebIE

Since the values on the WebIE data for each $\langle i_p, c_o \rangle \in \mathcal{R}_I$ are fractional in $[0, 1]$, we determined the lowest threshold with at least 0.80 precision and computed recall based on that, resulting in a recall of 0.136 at 0.80 precision for shopping, and a near-identical 0.139 for dining. This recall reflects the fraction of places with web pages, the fraction of offerings (products or dishes) mentioned on those pages, and the recall of the named entity recognition. We did not independently measure these other factors, as Web performance was merely a baseline. WALS on WebIE data was not able to show very significant improvement, and the results are not shown.

6.3.2. CS

The primary hypothesis of this paper is that the acquisition of class-level associations in $\mathcal{R}_C$ from the crowd is an effective way of rapidly jump-starting instance-level associations in $\mathcal{R}_I$. As described in Section 5, we acquired 25k class-level pairs from a paid crowd for shopping and 20k for dining, each with a score

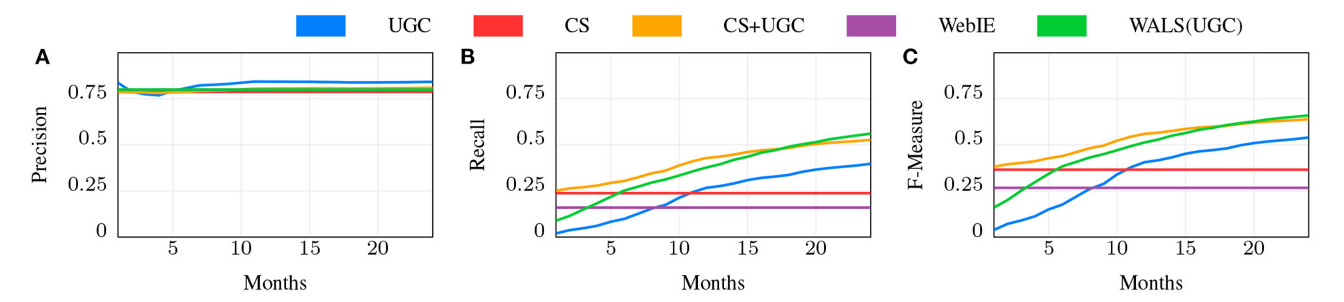


FIGURE 7 | Precision, Recall, and F-measure for different ways of predicting $\mathcal{R}_I$ for shopping.

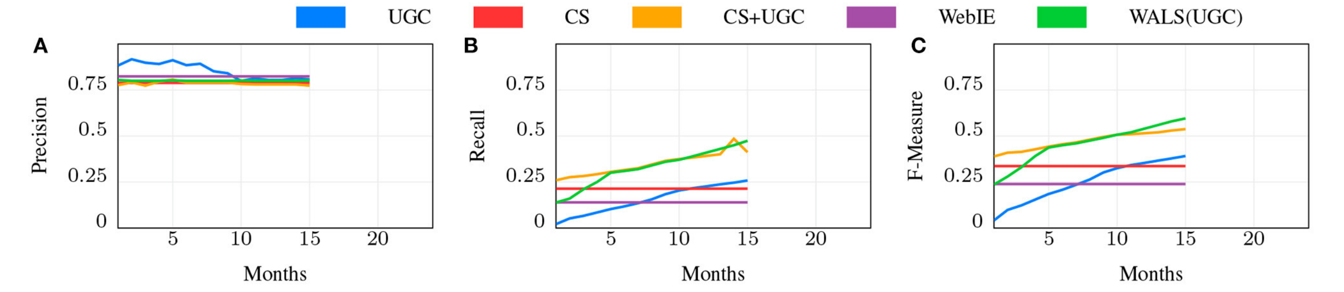


FIGURE 8 | Precision, Recall, and F-measure for different ways of predicting $\mathcal{R}_I$ for dining.

$w_{x,o}$ (see Section 5.5), and chose the following simple procedure to infer the instance level pairs:

$$w_{x,o} > 0.5 \wedge x \in \mathcal{C}_P \implies \langle x, o\rangle \in \mathcal{R}_C$$
$$\langle x, o\rangle \in \mathcal{R}_C \wedge \langle y, x\rangle \in \mathcal{R}_T \implies \langle y, o\rangle \in \mathcal{R}_I$$

In other words, for class level pair $\langle x, o\rangle$, if x is a place category, and the majority of raters ($w_{x,o} > 0.5$) answered that you can find o at places of that type, add a class-level edge to $\mathcal{G}$, and an instance-level edge to every instance of place category x.

We then measured the effectiveness of the CS by comparison of the inferred edges in $\mathcal{R}_I$ to the Gold set, achieving a recall of 0.238 with a precision of 0.788 for shopping, and 0.214 with a precision of 0.788 for dining. While this shows a distinct improvement over WebIE, of interest is the combination, which improves recall to 0.351—near perfect complementarity—while slightly losing precision at 0.782 (for simplicity we do not show this in **Figure 7** or **Figure 8**). The combination uses the WebIE or CS signal if the other is not present, and the CS signal if they are both present, since the CS data includes negatives and WebIE does not. (WALS inference was ineffective here; see below).

6.3.3. UGC

The UGC dataset grows over time as more users visit places and answer questions, while we treat the Web and CS data as constant (see above). We expect that, given enough time, UGC will overtake CS and WebIE in recall, so an important question is how much time the CS data is worth compared to UGC, and whether it continues to show value. In **Figures 7**, **8**, the blue line shows the precision, recall, and F1 score of the UGC data using the majority vote as the label, and the red line shows the CS performance, which, as noted above, doesn't change. In both shopping and dining, the UGC line crosses the CS line at around 11 months, indicating that CS is worth about 11 months of UGC collection in both domains.

6.3.4. WALS(UGC)

We populated the matrix $\mathbf{R}_{p,o}$ from UGC $w_{p,o}$ scores, factorized $\mathbf{R}$ using WALS, and measured the resulting dot-products against the Gold Standard dataset, shown in **Figures 7**, **8** in green. Since WALS produces real-valued predictions, we chose the 0.8 prec. threshold, the comparable precision of the CS and UGC methods, and measured the recall at that threshold with increasing UGC over time.

Note that some of the $\langle p, o\rangle$ pairs in the Gold set were in the training set, however the *labels* used in the training matrix may be different than Gold, making it a fair comparison. As in the previous experiments we broke the dataset into sets representing the first $n \in [1, 24]$ months of collected user responses. WALS clearly improves over UGC.

6.3.5. CS+UGC

While 11 months is the intersection point of the metric values for CS and UGC independently, the CS data is supposed to complement as well as jump-start the knowledge acquisition. We tested the role of CS over time using a simple "CS as default" combination, shown in **Figures 7**, **8** as CS+UGC, in which the UGC label is used if present, and the CS label is used if not. This line tracks the improvement in recall over time from UGC

TABLE 3 | Example gold standard $\mathcal{R}_l$ pairs.

Store	Category	loc	Product	Available
7-Eleven	Convenience store	US	Distilled water	FALSE
ALDI	Grocery store	US	Fruitcake	TRUE
AURORA MKT	Store	US	Men's Gloves	FALSE
Adams Pharmacy	Pharmacy	US	Kool aid	TRUE
Ag construcciones	Building materials	PY	Blinds	TRUE
Alanyurt Gıda	General store	TR	Razor blades	TRUE
Amorino	Ice cream shop	FR	Meat	FALSE
Barnes and Noble	Book store	US	Blankets	FALSE
Barstow Buick	Car dealer	US	Crown victoria	TRUE
Barstow Buick	Car dealer	US	Gears	TRUE
Bazar	bazar	BR	Mary kay	FALSE

collection, while jump starting at the recall of CS. This is a clear demonstration of our core research hypothesis.

Of particular interest is the comparison of WALS(UGC) with CS+UGC. The former does eventually surpass the latter for shopping after roughly 18 m (**Figure 7**), but the CS+UGC combination is a strong contender from an extremely simple method. This is again clear evidence of our core hypothesis. However, for dining the story is not so clear, as the WALS(UGC) very quickly reaches near-parity with CS+UGCafter only 5 months, and starts to improve over it in the 11th month of UGCcollection (**Figure 8**). The reason for this is not entirely clear, the dining matrix is smaller than shopping—the number of restaurants and the number of dishes are both smaller—meaning the same amount of data collection is a higher part of the total matrix. There may be something slightly easier about the restaurant problem as well—restaurant menus tend to be much smaller than the number of products sold in most stores. Perhaps most importantly, for the early part of gathering shopping UGC, we did not have the crowd sense data to guide the collecting, that was available after 6 months, whereas for dining we collected the crowd sense data first and it guided the collection from the start.

Other ways of filling the initial training matrix **R** by combining CS, UGC, and WebIE signals in various ways were tried but not included as they do not outperform WALS(UGC). Of note is that the CS signal does not work well with WALS, since it effectively does what WALS itself should do with enough data - filling in giant portions of the matrix with default values. Other machine learning approaches are certainly possible, indeed the launched *local search* system uses a deep neural network with many more features that are beyond the scope of this paper, and measured at the scale of the web. The three signals reported here are very signifant features of that system, and the full system improves significantly over search alone.

7. RELATED WORK

The core of this work is overcoming a *knowledge acquisition* bottleneck in acquiring data reflecting the availability of products at millions of brick and mortar stores worldwide. The approach of harnessing class-level knowledge to infer instance-level knowledge is based on a long standing idea in knowledge engineering, dating back at least as far as Minsky (1974). Other methods in the formal *knowledge representation* (KR) field have never scaled to the level necessary for our problem, nor have they considered the problem of how to acquire distributions instead of discrete facts.

Information Extraction (IE) methods perform knowledge acquisition of real-world entities from web text, and are discussed in Zang et al. (2013). Martínez-Rodríguez et al. (2020) present a survey of IE techniques for populating semantic structures, e.g., entity extraction and linking. In the context of shopping, research has mainly focused on product information extraction, e.g., crawling the Web for offers to maintain product catalogs as in Nguyen et al. (2011) and Qiu et al. (2015a), extracting product specifications and attributes as with Kannan et al. (2011), Qiu et al. (2015b), Zheng et al. (2018), and Wang et al. (2020), and IE methods for building product knowledge graphs such as Dong (2020) and Xu et al. (2020). Our paper defines a method for linking these already defined entities similar to Dong (2020), incorporating product and store taxonomy knowledge.

Knowledge Base Completion (KBC) is the problem of inferring missing entities and/or relations in an existing knowledge graph based on existing ones, such as via link prediction as in Bordes et al. (2013) or from a combination of sources such as Riedel et al. (2013). Our product × store category matrix (**Figure 4**) is inspired by the item-based collaborative filtering matrix introduced in recommender systems found in Sarwar et al. (2001) and Ekstrand et al. (2011), and we leverage a well-known collaborative filtering approach introduced in Koren et al. (2009) for KBC to demonstrate the additional power of inference on our knowledge graph.

We use a knowledge graph as the basic representation and, like most well known KGs, employ no general-purpose reasoning; hence, any inference we do must be defeasible. The most relevant KR area would be reasoning with defaults (e.g., Reiter, 1978; Lang, 2000), as our CS+UGC baseline mechanism for combining $\langle c_p, c_o \rangle$ with $\langle i_p, c_o \rangle$ pairs treats the first as a default and the second as an override. Beyond this simple combination strategy, which was first proposed in Quillian (1967), more sophisticated combinations of CS+UGC with other forms of evidence are done using optimizations from machine learning. The full *local shopping system* uses many signals, of which we've described only three, that are combined using a deep neural network that optimizes the prediction of observed labels for many billions of $\langle i_p, c_o \rangle$ pairs. While we exploit the taxonomies in $\mathcal{C}_P$ and especially $\mathcal{C}_O$ to optimize the selection of class-level pairs to acquire from workers as discussed in Lees et al. (2020), taxonomy-based reasoning was only used for negative associations. This negative inheritance was first observed by Deng et al. (2014).

IE and KBC techniques have advanced the state-of-the-art in capturing human knowledge in machine-readable form, but there is still the need for human curation and *crowdsourcing*. Important milestones for crowdsourcing knowledge acquisition at scale are Wikidata (Bollacker et al., 2008) and Freebase (Vrandečić and Krötzsch, 2014), where the crowd defines or curates real world entities and some relationships between them, typically driven by Wikipedia. With respect to KBC, Revenko et al. (2018) propose a method for crowdsourcing categorical common sense knowlegde from nonexperts for adding new relationships between nodes in the graph and ensuring consistencey with existing relations. However in all these sources, Taylor (2017) has pointed to the sparsity of graph edges expressing relations between the class-level nodes. Our work focuses directly on that problem by acquiring both class-level and instance level graph edges, and scaling the latter from the former.

The crowdsourcing approach we propose in this paper is grounded in the theoretical framework of Aroyo and Welty (2013) and Aroyo and Welty (2014), which breaks the constraints of typical methodologies for collecting ground truth, showing disagreement is a necessary characteristic of annotated data; when interpreted correctly, Dumitrache (2019) showed it can make evaluation of machine learning models more attuned to real-world data.

The immense body of research on common sense and crowdsourcing has directly influenced our work. The UGC and Crowd Sense tasks drew on our knowledge of Games-with-a-purpose such as Verbosity for collecting common sense facts (von Ahn et al., 2006), Common Consensus for gathering common sense goals (Lieberman et al., 2007), GECKA for common sense knowledge acquisition (Cambria et al., 2016), Concept Game for verifying common sense knowledge assertions (Herdagdelen and Baroni, 2010), the FACTory Game for facts verification (Lenat and Guha, 1989) and many others. Rodosthenous and Michael (2019) refer to common sense as "knowledge about the world" and propose a hybrid (machine and human tasks) workflow to gather general common sense knowledge rules.

Active learning investigates efficiency for acquisition and learning when acquiring training data for ML models. In essence, the early stages of KG acquisition strongly represent the exploration side of the *exploration vs. exploitation* tradeoff introduced by Bondu et al. (2010). ML models during exploration do not have enough knowledge of the space to be able to offer reliable judgements as to which items (in this case, $\langle i_p, c_o \rangle$ pairs) to acquire labels for. As noted in the Section 6.1, class-level pairs can serve as a guide for recognizing obvious $\langle i_p, c_o \rangle$ pairs that likely do not need labels, and conversely, high-disagreement pairs are very likely to have instances that do. Thus the $\langle c_p, c_o \rangle$ pairs can serve to stratify the $\langle i_p, c_o \rangle$ space, and make the job of active learning easier by narrowing down their targets. In Section 4 we discussed using these *possible* class-level pairs to guide sampling for UGC.

The problem of mining "interesting" negative statements from Wikidata was investigated in Karagiannis et al. (2019), Arnaout et al. (2020), and Arnaout et al. (2021), which in principle could be used to supplement our active learning strategies for selecting difficult training examples to improve the model. Specifically, these could be combined with the obvious (positive and negative) class-level pairs to find exceptions at individual stores, e.g., a grocery store that does not sell milk or that sells certain tools. Our approach would be slow to find such exceptions, since we don't ask users and would need other sources of evidence used by the larger production syste (e.g., a web page, a user review, etc.). *Peer-based detection*, which compares triples with other triples that share entities in the same category, is similar in spirit to collaborative filtering (CF) though they did not compare experimentally against a CF method such as WALS. *Pattern-based detection*, presented in Karagiannis et al. (2019) and Arnaout et al. (2020) seems better suited for mining (negative) trivia than for product availability, since it is unlikely many online users write about e.g., why supermarkets don't sell asphalt.

Perhaps the most similar crowdsourcing work to ours studies the problem of approximating aggregation queries presented in Trushkowsky et al. (2013), such as "How many restaurants in San Francisco serve scallops?" While this approach works well for estimating counts, clearly it does not scale for KBC.

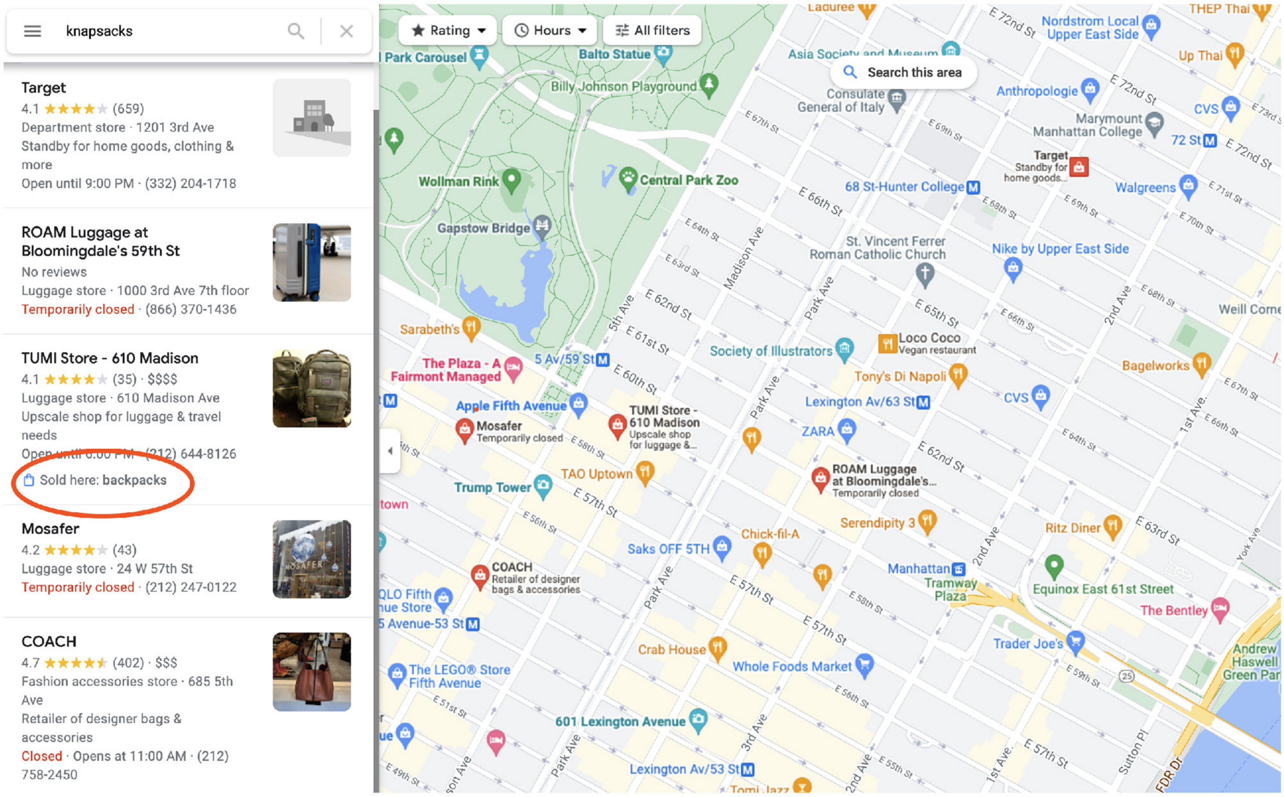
8. CONCLUSIONS

The CrowdSense approach was an integral part of a successful worldwide launch of local search results to queries for products or dishes, overlaid on Google Maps, as shown in **Figure 1**. Due to the complexity and scope of the deployed project, we focused on the real-world knowledge acquisition aspect of the work, and presented a few simplified experiments that demonstrate how the acquired class-level knowledge can be used for KBC at the instance level. These experiments may seem over-simplified, but they accurately capture the impact of the three-tiered crowdsourcing approach on the deployed product, in particular the rapid jump-start of the place-offering edges in the knowledge graph.

To achieve these results, we augmented an existing knowledge graph of most stores and restaurants on earth, their categories, dishes and a product taxonomy, by adding place to product and place to dish edges. We combined web-based information extraction (WebIE) and direct user observations collected over 2 years (UGC) with a novel collection of class-level $\langle$*store*, *offering*$\rangle$ pairs from the crowd (CS), which were inferred to the instance-level based on class membership. In 2 weeks of data collection we achieved a recall of 0.24 at 0.80 precision against gold standard instance-level labels for shopping, and 0.21 for dining. The class-level data for shopping combined with WebIE to achieve 0.35 recall, which was the recall of a WALS model with 18 months of UGC input. For dining the same combination also produced 0.34 recall, which was the WALs recall for 11 months of UGC. We conclude that the Crowd Sense approach uses human common sense knowledge to *rapidly jump start* the kind of generalization that ML systems are good at with a lot of data. This has implications for practical ML and Human Computation.

Our class-level crowdsourcing results show that the disagreement in categorical knowledge collected from the crowd can indicate the distribution of that knowledge at the instance level, rather than assuming the class-level associations are universally true: in other words, if 80% of raters say "Grocery stores sell oat milk," then $\sim$ 80% of grocery stores sell oat milk. These results held also for dishes at restaurants.

The taxonomy of products was used to guide the sampling of class-level pairs in a way that helped us address the sparsity of the $\mathcal{C}_P \times \mathcal{C}_O$ space, and only the *negative* class-level attributes were accurate when inferred to more specific categories, as in Deng et al. (2014), as opposed to the more traditional view that positive attributes are "inherited."



knapsacks

Target
4.1 ★★★★ (659)
Department store · 1201 3rd Ave
Standby for home goods, clothing & more
Open until 9:00 PM · (332) 204-1718

ROAM Luggage at Bloomingdale's 59th St
No reviews
Luggage store · 1000 3rd Ave 7th floor
Temporarily closed · (866) 370-1436

TUMI Store - 610 Madison
4.1 ★★★★ (35) · $$$$
Luggage store · 610 Madison Ave
Upscale shop for luggage & travel needs
Open until 6:00 PM · (212) 644-8126
Sold here: backpacks

Mosafer
4.2 ★★★★ (43)
Luggage store · 24 W 57th St
Temporarily closed · (212) 247-0122

COACH
4.7 ★★★★★ (402) · $$$
Fashion accessories store · 685 5th Ave
Retailer of designer bags & accessories
Closed · Opens at 11:00 AM · (212) 758-2450

FIGURE 9 | CrowdSense search results in NYC for knapsacks.

We found the categorical pairs which were rapidly acquired were extremely useful in guiding the collection of instance-level labels, since we did not have to ask users about obviously available or unavailable products—this has implications for active learning, and held also for dining.

We expected the class-level ratings we acquired from a small, international, pool of paid raters, to show bias toward ratings coming from the same country of a restaurant. In other words, we expected class-level ratings from Indian raters to have lower error for restaurants in India than class-level ratings from raters in other countries. This turned out to only be true for Japan, and for all other countries it was the opposite. This may tell us something about the way the place categories model the real world, more investigation is required.

We believe Crowd Sense is a general technique for knowledge acquisition that can provide a rapid jump-start to the process by acquiring more general, common-sense defaults as a first step, while more precise but time-consuming acquisition (i.e., at the instance level) proceeds over time. We have shown that the original local shopping idea, first presented in Welty et al. (2021), can generalize to other establishment domains with similar gains, in this case dining, and we have considered many other bipartite problems that meet the basic requirement that there is a strong, common-sense understanding of the relation at the categorical level, for example:

- *Dish contains ingredient.* Dishes have associated recipes and a strong notion of taxonomy[8], and many ingredient associations are ridiculous at a class level, such as Apple Pie and Curry.
- *Cuisine includes dish.* Dishes are also associated with cuisines, a pairing that could be useful for recipe datasets, and understanding menus. Many cuisines are regional, introducing a different kind of partial order (containment rather than generalization, see Guarino and Welty, 2009) on one side of the bipartite relation.
- *Wildlife inhabiting a region.* Several NGOs track wildlife populations through remote cameras and citizen science collection of photos, and identify animals using automatic methods.[9] Such methods would benefit from large scale understanding of obvious negatives (tigers are not found in Africa). Like cuisines, this involves treating locations as a partial order based on containment, and the Linnaean taxonomy for animals is well established.
- *Animal has body part.* In the early days of AI, much ink was spilled on modeling defaults and exceptions such as "Elephants have trunks" and "Humans have two legs." This work was summarized nicely in Brachman (1985). Modern AI systems do not use this information and rely on the

[8] e.g., https://www.wikidata.org/wiki/Wikidata:WikiProject_Food/Taxonomy.
[9] Examples include wildlifeinsights.org and inaturalist.org.

formation of embeddings that bely human understanding, but such systems have been shown in Aroyo and Paritosh (2021) to make "silly" categorical mistakes. An approach that forces large models to form meaningful intermediate representations such as parts of the body, as described by Hinton (2021), could avoid silly mistakes with this form of common sense curation.

- *Company owns patent.* Finding patents is a difficult search task that continues to be a focus of AI systems. While these systems do not generally lack data, they do often suffer from silly mistakes, as image understanding systems do, which reflect a lack of common sense. Adding categorical associations such as, "Tech companies do not own pharmaceutical patents" would eliminate some of these mistakes.

To see CrowdSense at work, type the name of a product or dish into Google Maps (or Google Search). Results that say "Sold here: *product*" come from the data we published (see **Figure 9**, as opposed to "In stock" (merchant feeds) and "Webpage says." Anyone with a Google account can participate in UGC (user generated content) acquisition. Users with location tracking turned on (so that maps knows what places the user has visited[10]) can navigate to the "contribute" tab that allows them to rate and leave reviews, as well as review facts and answer the yes/no questions regarding locations they have visited.

[10] See https://support.google.com/local-guides/answer/6225846.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s. The categories of places and products are available and included in the article, and the rest of the data discussed in this paper is not, as it is proprietary data that drives Google Maps.

AUTHOR CONTRIBUTIONS

CW led the project. CW and FK wrote most of the article. All authors contributed experimental results and background research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.830299/full#supplementary-material

REFERENCES

Arnaout, H., Razniewski, S., and Weikum, G. (2020). "Enriching knowledge bases with interesting negative statements," in *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*, eds D. Das, H. Hajishirzi, A. McCallum, and S. Singh.

Arnaout, H., Razniewski, S., Weikum, G., and Pan, J. Z. (2021). "Negative knowledge for open-world wikidata," in *Companion of The Web Conference 2021, Virtual Event/Ljubljana, Slovenia, April 19–23, 2021*, eds J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and L. Zia (ACM/IW3C2), 544–551.

Aroyo, L., and Paritosh, P. (2021). *Uncovering Unknown Unknowns in Machine Learning.* Google AI Blog. Available online at: https://ai.googleblog.com/2021/02/uncovering-unknown-unknowns-in-machine.html

Aroyo, L., and Welty, C. (2013). "Crowd truth: harnessing disagreement in crowdsourcing a relation extraction gold standard," in *Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13*.

Aroyo, L., and Welty, C. (2014). The three sides of crowdtruth. *Hum. Comput.* 1, 31–44. doi: 10.15346/hc.v1i1.3

Aroyo, L., and Welty, C. (2015). Truth is a lie: Crowd Truth and the seven myths of human annotation. *AI Mag.* 36, 15–24. doi: 10.1609/aimag.v36i1.2564

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of SIGMOD International Conference on Management of Data* (New York, NY: Association for Computing Machinery), 1247–1250.

Bondu, A., Lemaire, V., and Boullé, M. (2010). "Exploration vs. exploitation in active learning : a bayesian approach," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*.

Bordes, A., Usunier, N., García-Durán, A., Weston, J., and Yakhnenko, O. (2013). "Translating embeddings for modeling multi-relational data," in *NIPS 2013*, eds C. J. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, 2787–2795.

Brachman, R. J. (1985). I lied about the trees, or, defaults and definitions in knowledge representation. *AI Mag.* 6, 80.

Cambria, E., Nguyen, T. V., Cheng, B., Kwok, K., and Sepulveda, J. (2016). "Gecka3d: A 3d game engine for commonsense knowledge acquisition," in *CoRR-2016*.

Deng, J., Russakovsky, O., Krause, J., Bernstein, M., Berg, A., and Fei-Fei, L. (2014). "Scalable multi-label annotation," in *Proceedings of CHI 2014* (New York, NY: Association for Computing Machinery).

Dong, X. L. (2020). "Autoknow: self-driving knowledge collection for products of thousands of types," in *KDD '20: Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 2724–2734.

Dumitrache, A. (2019). *Truth in Disagreement: Crowdsourcing Labeled Data for Natural Language Processing* (Ph.D. thesis), VU Amsterdam.

Ekstrand, M. D., Riedl, J. T., and Konstan, J. A. (2011). *Collaborative Filtering Recommender Systems*. Norwell, MA: Now Publishers Inc. doi: 10.1561/9781601984432

Guarino, N., and Welty, C. (2009). "An overview of OntoClean," in *Handbook on Ontologies. International Handbooks on Information Systems* (Berlin; Heidelberg: Springer), 201–220.

Herdagdelen, A., and Baroni, M. (2010). "The concept game: Better commonsense knowledge extraction by combining text mining and a game with a purpose," in *AAAI Fall Symposium: Commonsense Knowledge*.

Hinton, G. E. (2021). How to represent part-whole hierarchies in a neural network. *CoRR, abs*/2102. *12627*.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2020). A survey on knowledge graphs: Representation, acquisition and applications. *CoRR*, abs/*2002.00388*.

Kannan, A., Givoni, I. E., Agrawal, R., and Fuxman, A. (2011). "Matching unstructured product offers to structured product specifications," *in KDD-2011*, 404–412.

Karagiannis, G., Trummer, I., Jo, S., Khandelwal, S., Wang, X., and Yu, C. (2019). Mining an "anti-knowledge base" from wikipedia updates with applications to fact checking and beyond. *Proc. VLDB Endow*. 13, 561–573. doi: 10.14778/3372716.3372727

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* 42, 30–37. doi: 10.1109/MC.2009.263

Lagos, N., Ait-Mokhtar, S., and Calapodescu, I. (2020). "Point-of-interest semantic tag completion in a global crowdsourced search-and-discovery database," in *ECAI 2020, volume 325 of Frontiers in Artificial Intelligence and Applications*, eds G. D. Giacomo, A. Catalá, B. Dilkina, M.

Milano, S. Barro, A. Bugarín, and J. Lang (Amsterdam: IOS Press), 2993–3000.

Lang, J. (2000). *Possibilistic Logic: Complexity and Algorithms*. Dordrecht: Springer Netherlands.

Lees, A. W., Welty, C., Korycki, J., Carthy, S. M., and Zhao, S. (2020). "Embedding semantic taxonomies," in *CoLing 2020*.

Lenat, D. B., and Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc.

Lieberman, H., Smith, D., and Teeters, A. (2007). "Common consensus: a web-based game for collecting commonsense goals," in *IUI-2007*.

Martínez-Rodríguez, J. L., Hogan, A., and López-Arévalo, I. (2020). Information extraction meets the semantic web: a survey. *Semantic Web* 11:255–335. doi: 10.3233/SW-180333

McNamee, P., and Dang, H. T. (2009). "Overview of the tac 2009 knowledge base population track," in *Text Analysis Conference (TAC-2009)*.

Minsky, M. (1974). *A Framework for Representing Knowledge*. Technical report, MIT.

Nguyen, H., Fuxman, A., Paparizos, S., Freire, J., and Agrawal, R. (2011). Synthesizing products for online catalogs. *Proc. VLDB Endow*. 4, 409–418. doi: 10.14778/1988776.1988777

Noy, N., Rector, A., Hayes, P., and Welty, C. (2006). *Defining N-ary Relations on the Semantic Web*. Available online at: http://www.w3.org/TR/swbp-n-aryRelations

Qiu, D., Barbosa, L., Dong, L. X., Shen, Y., and Srivastava, D. (2015a). "Dexter: large-scale discovery and extraction of product specifications on the web," in *PVLDB*, 2194–2205.

Qiu, D., Barbosa, L., Dong, X. L., Shen, Y., and Srivastava, D. (2015b). DEXTER: large-scale discovery and extraction of product specifications on the web. *Proc. VLDB Endow*. 8, 2194–2205. doi: 10.14778/2831360.2831372

Quillian, M. R. (1967). Word concepts: a theory and simulation of some basic semantic capabilities. *Behav. Sci*. 12, 410–430. doi: 10.1002/bs.3830120511

Reiter, R. (1978). "On reasoning by default," in *Readings in Knowledge Representation*, eds R. J. Brachman and H. Levesque (San Francisco, CA: Morgan Kaufmann Publishers Inc.).

Revenko, A., Sabou, M., Ahmeti, A., and Schauer, M. (2018). "Crowd-sourced knowledge graph extension: a belief revision based approach," in *Proceedings of the HCOMP 2018 Works in Progress and Demonstration Papers Track, volume 2173 of CEUR Workshop Proceedings*, eds A. Bozzon and M. Venanzi (CEUR-WS.org).

Riedel, S., Yao, L., Marlin, B. M., and McCallum, A. (2013). "Relation extraction with matrix factorization and universal schemas," in *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (HLT-NAACL).

Rodosthenous, C., and Michael, L. (2019). "A platform for commonsense knowledge acquisition using crowdsourcing," in *Proceedings of the enetCollect WG3 and WG5 Meeting 2018, Vol. 2390*, 25–30.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., and Aroyo, L. M. (2021). *"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes ai.*

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web* (New York, NY: Association for Computing Machinery), 285–295.

Shortliffe, E., and Buchanan, B. (1975). A model of inexact reasoning in medicine. *Math. Biosci*. 23, 351–379. doi: 10.1016/0025-5564(75)90047-4

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Taylor, J. (2017). *Iswc 2017 Keynote: Applied Semantics: Beyond the ca||Talog*. Available online at: https://iswc2017.ai.wu.ac.at/program/keynotes/keynote-taylor/.

Trushkowsky, B., Kraska, T., Franklin, M. J., and Sarkar, P. (2013). "Crowdsourced enumeration queries," in *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8–12, 2013* (IEEE Computer Society), 673–684.

von Ahn, L., Kedia, M., and Blum, M. (2006). "Verbosity: a game for collecting common-sense facts," in *Proceedings Conference on Human Factors in Computing Systems, CHI 2006* (New York, NY: ACM), 75–78.

Vrandečić, D., and Krötzsch, M. (2014). Wikidata: a free collaborative knowledge base. *Commun, ACM* 57, 78–85. doi: 10.1145/2629489

Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., et al. (2020). "Learning to extract attribute value from product via question answering: a multi-task approach," in *KDD-20*.

Welty, C., Aroyo, L., Korn, F., McCarthy, S., and Zhao, S. (2021). "Rapid instance-level knowledge acquisition for google maps from class-level common sense," in *Proceedings of HCOMP-2021. AAAI*.

Welty, C., Barker, K., Aroyo, L., and Arora, S. (2012). "Query driven hypothesis generation for answering queries over nlp graphs," in *The Semantic Web-ISWC 2012* (Berlin; Heidelberg: Springer), 228–242.

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., and Achan, K. (2020). "Product knowledge graph embedding for e-commerce," in *Proceedings of Conference on Web Search and Data Mining, WSDM '20*, 672–680.

Zang, L., Cao, C., Cao, Y., Wu, Y., and Cao, C. (2013). A survey of commonsense knowledge acquisition. *J. Comput. Sci. Technol*. 28, 689–719. doi: 10.1007/s11390-013-1369-6

Zheng, G., Mukherjee, S., Dong, X. L., and Li, F. (2018). "Opentag: Open attribute value extraction from product profiles," in *Proceedings of KDD '18* (New York, NY: Association for Computing Machinery), 1049–1058.

Conflict of Interest: All authors were employed by Google Research.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/frai.2022.798892

Utility of Crowdsourced User Experiments for Measuring the Central Tendency of User Performance: A Case of Error-Rate Model Evaluation in a Pointing Task

*Shota Yamanaka**

Yahoo! JAPAN Research, Yahoo Japan Corporation, Tokyo, Japan

OPEN ACCESS

Edited by:
Ujwal Gadiraju,
Delft University of Technology, Netherlands

Reviewed by:
Evangelos Niforatos,
Delft University of Technology, Netherlands
Helena Macedo Reis,
Federal University of Paraná, Brazil

Correspondence:
Shota Yamanaka
syamanak@yahoo-corp.jp

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *20 October 2021*
Accepted: *21 February 2022*
Published: *17 March 2022*

Citation:
Yamanaka S (2022) Utility of Crowdsourced User Experiments for Measuring the Central Tendency of User Performance: A Case of Error-Rate Model Evaluation in a Pointing Task.
Front. Artif. Intell. 5:798892.
doi: 10.3389/frai.2022.798892

The usage of crowdsourcing to recruit numerous participants has been recognized as beneficial in the human-computer interaction (HCI) field, such as for designing user interfaces and validating user performance models. In this work, we investigate its effectiveness for evaluating an error-rate prediction model in target pointing tasks. In contrast to models for operational times, a clicking error (i.e., missing a target) occurs by chance at a certain probability, e.g., 5%. Therefore, in traditional laboratory-based experiments, a lot of repetitions are needed to measure the central tendency of error rates. We hypothesize that recruiting many workers would enable us to keep the number of repetitions per worker much smaller. We collected data from 384 workers and found that existing models on operational time and error rate showed good fits (both $R^2 > 0.95$). A simulation where we changed the number of participants N_P and the number of repetitions N_{repeat} showed that the time prediction model was robust against small N_P and N_{repeat}, although the error-rate model fitness was considerably degraded. These findings empirically demonstrate a new utility of crowdsourced user experiments for collecting numerous participants, which should be of great use to HCI researchers for their evaluation studies.

Keywords: crowdsourcing, graphical user interface, Fitts'law, user performance models, error-rate prediction

1. INTRODUCTION

In the field of human-computer interaction (HCI), a major topic is to measure the time needed to complete a given task for (e.g.,) evaluating novel systems and techniques. Examples include measuring a text-entry time (Banovic et al., 2019; Cui et al., 2020), a time to learn a new keyboard layout (Jokinen et al., 2017), and a menu-selection time (Bailly et al., 2016). In these studies, generally, laboratory-based user experiments have been conducted. That is, researchers recruit ten to 20 students from a local university and ask them to use a specified apparatus to perform a task in a silent room. However, researchers are aware of the risk of conducting a user experiment with a small sample size; e.g., the statistical power is weak (Caine, 2016). Therefore, using crowdsourcing services to recruit numerous participants has recently become more common, particularly for user experiments on graphical user interfaces (GUIs), e.g., (Komarov et al., 2013; Matejka et al., 2016; Findlater et al., 2017; Yamanaka et al., 2019; Cockburn et al., 2020).

There are two representative topics for research involving GUIs. The first is designing better GUIs or interaction techniques. In typical user experiments, researchers would like to compare a new GUI or technique with a baseline to demonstrate that a proposed one is statistically better. For this purpose, recruiting numerous participants is effective in finding statistical differences.

The other topic involving GUI experiments is deriving user performance models and empirically validating them. Conventionally, there are two representative metrics for GUI operations to be modeled: time and error rate (Wobbrock et al., 2008). A well-known model in HCI is Fitts' law (Fitts, 1954) to predict the operational time for target pointing tasks, or referred to as *Fitts's law* in some papers (MacKenzie, 2002). In lab-based user experiments to evaluate the model fitness in terms of R^2, university student participants typically join a study and are asked to point to a target repeatedly. For example, researchers set three target distances and three target sizes (i.e., nine task conditions in total), and the participants repeatedly click a target 15 times for each task condition. The average time for these 15 clicks is recorded as the final score for a participant (Soukoreff and MacKenzie, 2004).

In addition to operation times, the importance of predicting how accurately users can perform a task has recently been emphasized (Bi and Zhai, 2016; Huang et al., 2018, 2020; Park and Lee, 2018; Yamanaka et al., 2020; Do et al., 2021). In contrast to measuring the target-pointing times, where the time to click a target can be measured in every trial, the error rate is computed after repeatedly performing a single task condition (15 trials in the above-mentioned case). For example, if a participant misses a target in one trial, the error rate is recorded as $1/15 \times 100\% = 6.67\%$; if there are ten participants, one miss corresponds to 0.667% in the end. Because errors can occur by chance, evaluating error-rate models often requires more data (repetitions) for each task condition to measure the central tendency of the error rate. To evaluate the model's prediction accuracy more precisely, researchers have asked participants to perform more repetitions, as it is often difficult to collect numerous participants for lab-based experiments. For example, a previous study on touch-based error-rate models set 40 repetitions for each task condition collected from 12 participants. In this case, one miss corresponded to a 0.208% error rate (Yamanaka and Usuba, 2020).

However, for crowdsourced user experiments with GUIs, researchers cannot set a large number of repetitions per task condition. To enable crowdworkers to concentrate on a given task, it is recommended to set short task completion times, as workers switch to other tasks every 5 min on average (Gould et al., 2016). Hence, forcing a routine GUI operation task that takes, e.g., 40 min (Huang et al., 2018) or 1 h (Park and Lee, 2018; Yamanaka et al., 2020) would be harmful in terms of accurate measurement of the error rates. This could be considered a disadvantage of crowdsourced GUI study. An alternative to increasing the number of repetitions per task condition is simply to recruit more workers. This would enable the error rates to be measured more precisely, which would lead to a good prediction accuracy by the error-rate model (our research hypothesis). Even if the number of repetitions is only ten, utilizing 300 workers would mean that one miss corresponds to 0.033%. This is much more precise than the above-mentioned examples with error rates such as 0.208%.

However, there are several crowdsourcing-specific uncertainties that might affect the user performance results. For example, crowdworkers use different mice, displays, operating systems, cursor speed configurations, and so on; these factors significantly affect the target pointing performance in terms of both time and accuracy (MacKenzie et al., 2001; Casiez and Roussel, 2011). In addition, while studies have shown that the performance model on time (Fitts' law) is valid for crowdsourced data, crowdworkers tend to be more inaccurate than lab-based participants in target pointing tasks (Komarov et al., 2013), where error rates approximately two times higher or more have been observed (Findlater et al., 2017). Therefore, we would avoid claiming that user-performance models validated in crowdsourced studies are always applicable to lab-based controlled experiments. Also, it is not reasonable to interpret that the results such as error rates and operational times are directly comparable with lab-based participants.

Nevertheless, if an error-rate model we test exhibits a good fit (e.g., $R^2 > 0.9$), HCI researchers would have access to a powerful tool, crowdsourcing, to evaluate their newly proposed error-rate prediction models. Such a result stands to expand the application range of crowdsourcing in HCI; this motivated us to conduct this work. Our contributions are as follows.

- We conducted a crowdsourced mouse-pointing experiment following the Fitts' law paradigm. In total, we recorded 92,160 clicks performed by 384 crowd workers. Our error-rate model showed a good fit with $R^2 = 0.9581$, and cross-validation confirmed that the model can predict new (unknown) task conditions, too. This is the first study that demonstrates a GUI error-rate model holding to crowdsourced user data.
- We simulated how the number of participants N_P and the number of repetitions per task condition N_{repeat} affected the model fitness. We randomly sampled a limited portion of the entire workers (N_P from 10 to 320), and while each worker performed ten trials per task condition, we used only the data for the first N_{repeat} trials (from 2 to 10). After testing the model fitness over 1,000 iterations, we found that increasing N_P improved the prediction accuracy as well as increasing N_{repeat} could. The effect of N_P and N_{repeat} on the fitness was more clearly observed for the error-rate model than the time model, which suggests that crowdsourcing services are more suitable for evaluating novel error-rate models.

This article is an extended version of our previous work presented at the AAAI HCOMP 2021 conference (Yamanaka, 2021b). The points of difference are mainly twofold. First, to analyze the empirical data in more detail, this article newly shows figures that visualize statistically significant differences for the main and interaction effects of independent variables on the outcomes (operational time, click-point variability, and error rate) (see **Figures 3**, **5**, **7**). Second, we re-ran the simulation in which the random-sampling was repeatedly performed over 1,000 iterations, while in the conference-paper version we did

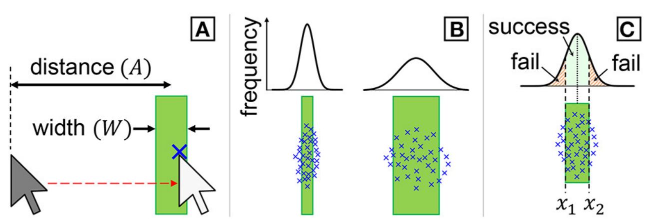


FIGURE 1 | (A) We use the Fitts' law paradigm in which users point to a vertically long target. A clicked position is illustrated with an "x" mark. **(B)** It has been assumed that the click positions recorded in many trials distribute normally, and its variability would increase with the target width. **(C)** An error rate is computed based on the probability where a click falls outside the target.

it over 100 iterations. This larger number of iterations gives us more reliable, less noisy data. We also newly added the standard deviation *SD* values of the model fitness for the 1,000 iterations for the sake of completeness (see **Figure 9**). Several discussions on these new results, such as comparisons with previous studies regarding model fitness, are also added in this revision.

2. RELATED WORK

2.1. Time Prediction for Pointing Tasks

For comparing the sensitivity of time and error-rate prediction models against N_P and N_{repeat}, we examine a robust time-prediction model, called Fitts' law (Fitts, 1954). According to this model, the time for the first click, or movement time *MT*, to point to a target is linearly related to the index of difficulty *ID* measured in bits:

$$MT = a + b \cdot ID = a + b \cdot \left(\frac{A}{W} + 1\right), \tag{1}$$

where a and b are empirical regression constants, A is the target distance (or amplitude), and W is its width (see **Figure 1A**). There are numerous formulae for calculating the *ID*, such as using a square root instead of the logarithm or using the effective target width (Plamondon and Alimi, 1997), but previous studies have shown that Equation 1 yields excellent model fitness (Soukoreff and MacKenzie, 2004). Using this Fitts' law, researchers can measure *MT*s for several $\{A, W\}$ conditions, regress the data to compute a and b, and then predict the *MT* for a new $\{A, W\}$ condition by applying the parameters of $\{a, b, A, W\}$ to Equation 1.

2.2. Error-Rate Prediction for Pointing Tasks

Researchers have also tried to derive models to predict the error rate *ER* (Meyer et al., 1988; Wobbrock et al., 2008; Park and Lee, 2018). In practice, the *ER* should increase as participants move faster, and vice versa (Zhai et al., 2004; Batmaz and Stuerzlinger, 2021). In typical target pointing experiments, participants are instructed to "point to the target as quickly and accurately as possible," which is intended to balance the speed and carefulness to decrease both *MT* and *ER* (MacKenzie, 1992; Soukoreff and MacKenzie, 2004).

In pointing tasks, as the target size decreases, users have to aim for the target more carefully to avoid misses. Accordingly, the spread of click positions should be smaller. If researchers conduct a pointing experiment following a typical Fitts' law methodology, in which two vertically long targets are used and participants perform left-right cursor movements, the click positions would follow a normal distribution (**Figure 1B**) (Crossman, 1956; MacKenzie, 1992). Formally speaking, a click point is a random variable X following normal distribution: $X \sim N(\mu, \sigma^2)$, where μ and σ are the mean and standard deviation of the click positions on the x-axis, respectively. The click point variability σ is assumed to proportionally relate to the target width, or to need an intercept, i.e., linear relationship (Bi and Zhai, 2016; Yu et al., 2019; Yamanaka and Usuba, 2020):

$$\sigma = c + d \cdot W, \tag{2}$$

where c and d are regression constants. The probability density function for a normal distribution, $f(x)$, is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{3}$$

If we define the target center as located at $x = 0$ with the target boundary ranging from x_1 to x_2 (**Figure 1C**), the predicted probability for where the click point X falls on the target, $P(x_1 \leq X \leq x_2)$, is

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= \int_{x_1}^{x_2} f(x)dx \\ &= \frac{1}{2}\left[\text{erf}\left(\frac{x_2-\mu}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{x_1-\mu}{\sigma\sqrt{2}}\right)\right], \end{aligned} \tag{4}$$

where erf(·) is the Gauss error function:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt. \tag{5}$$

Previous studies have shown that the mean click point is located close to the target center ($\mu \approx 0$), and σ is not significantly affected by the target distance A (MacKenzie, 1992; Bi and Zhai, 2016; Yamanaka and Usuba, 2020). Given the target width W, Equation 4 can be simplified and the *ER* is predicted as

$$\begin{aligned} ER &= 1 - P\left(-\frac{W}{2} \leq X \leq \frac{W}{2}\right) = 1 \\ &- \frac{1}{2}\left[\text{erf}\left(\frac{W/2}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{-W/2}{\sigma\sqrt{2}}\right)\right] = 1 - \text{erf}\left(\frac{W}{2\sqrt{2}\sigma}\right). \end{aligned} \tag{6}$$

Similarly to the way Fitts' law is used, researchers measure σ for several $\{A, W\}$ conditions, regress the data to compute c and d in Equation 2, and then predict the σ for a new $\{A, W\}$ condition. In this way (i.e., using the predicted σ based on a new W), we can predict the *ER* with Equation 6 for a new task condition. While there are similar but more complicated versions of this model tuned for pointing tasks in virtual reality systems (Yu et al., 2019) and touchscreens (Bi and Zhai, 2016), to our knowledge, there has been no report on the evaluation of this model for the most fundamental computer environment, i.e., PCs with mice.

2.3. Crowdsourced Studies on User Performance and Model Evaluation for GUIs

For target pointing tasks in PC environments, Komarov et al. (2013) found that crowdsourced and lab-based experiments led to the same conclusions on user performance, such as that a novel facilitation technique called *Bubble Cursor* (Grossman and Balakrishnan, 2005) reduced the *MT* compared with the baseline point-and-click method. Yamanaka et al. (2019) tested the effects of target margins on touch-pointing performance using smartphones and reported that the same effects were consistently found in crowdsourced and lab-based experiments, e.g., wider margins significantly decreased the *MT* but increased the *ER*. Findlater et al. (2017) showed that crowdworkers had significantly shorter *MT*s and higher *ER*s than lab-based participants in both mouse- and touch-pointing tasks. Thus, they concluded that crowdworkers were more biased towards speed than accuracy when instructed to "operate as rapidly and accurately as possible."

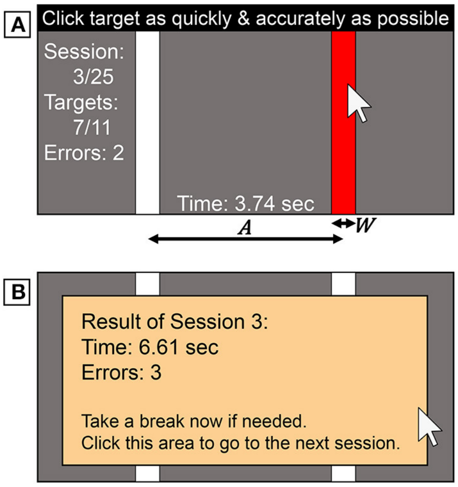
Regarding Fitts' law fitness, Findlater et al. reported that crowdworkers had average values of Pearson's $r = 0.926$ with mice and $r = 0.898$ with touchscreens (Findlater et al., 2017). Schwab et al. (2019) conducted crowdsourced scrolling tasks and found that Fitts' law held with $R^2 = 0.983$ and 0.972 for the desktop and mobile cases, respectively (note that scrolling operations follow Fitts' law well Zhao et al., 2014). Overall, these reports suggest that Fitts' law is valid for crowdsourced data regardless of the input device. It is unclear, however, how the N_P affects model fitness, because these studies used the entire workers' data for model fitting.

The only article that tested the effect of N_P on the fitness of user-performance models is a recent work by Yamanaka (2021a). He tested modified versions of Fitts' law to predict *MT*s in a rectangular-target pointing task. The conclusion was that, although he changed N_P from 5 to 100, the best-fit model did not change. However, because he used all N_{repeat} clicks, increasing N_P always increased the total data points to be analyzed, and thus the contributions of N_P and N_{repeat} could not be analyzed separately. We further analyze this point in our simulation.

In summary, there is a consensus that a time prediction model for pointing tasks (Fitts' law) shows a good fit for crowdsourced data. However, *ER* data have typically been reported as secondary results when measuring user performance in these studies. At least, no studies on evaluating *ER* prediction models have been reported so far. If we can demonstrate the potential of crowdsourced *ER* model evaluation, at least for one example task (target pointing in a PC environment), it will motivate future researchers to investigate novel *ER* models with less recruitment effort, more diversity of participants, and less time-consuming data collection. This will directly benefit the contribution of crowdsourcing to the HCI field.

3. USER EXPERIMENT

We conducted a traditional cyclic target-pointing experiment on the *Yahoo! Crowdsourcing* platform (https://crowdsourcing.yahoo.co.jp). Our affiliation's IRB-equivalent research ethics team approved this study. The experimental system was developed with the `Hot Soup Processor` programming language. The crowdworkers were asked to download and run an executable file to perform the experimental task.



Click target as quickly & accurately as possible

Session: 3/25
Targets: 7/11
Errors: 2

Time: 3.74 sec

Result of Session 3:
Time: 6.61 sec
Errors: 3

Take a break now if needed.
Click this area to go to the next session.

FIGURE 2 | Task stimuli used in the experiment. **(A)** Participants clicked alternately on each target when it was red. **(B)** At the end of a session, the results and a message to take a break were shown.

3.1. Task, Design, and Procedure

In the task window (1200 × 700 pixels), two vertically long targets were displayed (**Figure 2A**). If the participants clicked the target, the red target and white non-target rectangles switched colors, and they successively performed this action back and forth. If the participants missed the target, it flashed yellow, and they had to keep trying until successfully clicking it. We did not give auditory feedback for success or failure, as not all the participants would have been able to hear sound during the task. A *session* consisted of 11 cyclic clicks with a fixed $A \times W$ condition. The first click acted as a starting signal as we could not measure the *MT*, and thus the remaining ten trials for each session were used for data analysis. After completing a session, the participant saw the results and a message to take a break (**Figure 2B**).

The experiment was a 3×8 within-subjects repeated-measures design with the following independent variables and levels: three target distances ($A = 300$, 460, and 630 pixels) and eight widths ($W = 8$, 12, 18, 26, 36, 48, 62, and 78 pixels). These values were selected so that the values of *ID* ranged widely from 2.28 to 6.32 bits, which sufficiently covered easy to hard conditions according to a survey (Soukoreff and MacKenzie, 2004). Each participant completed 24 ($= 3_A \times 8_W$) sessions. The order of the 24 conditions was randomized. Before the first session, to allow the participants to get used to the task, they performed a practice session under a condition with $A = 400$ and $W = 31$ pixels, i.e., parameters that were not used in the actual 24 data-collection sessions. This experimental design was tuned with reference to the author's pilot study; without having a break, the task completion time was 3 min 40 s on average, which meets the recommendation for crowdsourced user experiments (Gould et al., 2016).

The *MT* was measured from when the previous target was successfully clicked to when the next click was performed regardless of the success or failure (MacKenzie, 1992; Soukoreff and MacKenzie, 2004). Trials in which we observed one or more clicks outside the target were flagged as an error. The first left target acted as a starting button, and the remaining ten trials' data were measured to compute *MT*, σ, and *ER*. After finishing all sessions, the participants completed a questionnaire on their age (numeric), gender (free-form to allow non-binary or arbitrary answers), handedness (left or right), Windows version (free-form), input device (free-form), and history of PC use (numeric in years).

3.2. Participants and Recruitment

We recruited workers who used Windows Vista or a later version to run our system. We requested no specific PC skills, as we did not wish to limit our collection to only high-performance workers' data. Also, we did not use any *a-priori* filtering options, such as the approval-rate threshold, which require additional cost for the crowdsourcing service. We made this decision because, if our hypothesis is supported with a less costly method, it would be more beneficial for future research to recruit many more participants with low cost for obtaining the central tendency of error rates. Still, clear outlier workers who seemed not to follow our instructions (such as performing the task too slowly) were removed when we analyzed the data. As we show later in the simulation analysis, this decision was not problematic because Fitts' law held well even if we analyzed only ten workers' data over 1,000 iterations (i.e., they exhibited typical rapid-and-accurate pointing behavior).

On the recruitment page, we asked the workers to use a mouse if possible. We made this request because, in our simulation analysis, we randomly selected a certain number of participants (e.g., $N_P = 10$) to examine if the model fitness was good or poor. If these workers used different devices (e.g., six mice, two touchpads, and two trackballs), we might have wondered if a poor model fit was due to the device differences. Nevertheless, to avoid a possible false report in which all workers might answer they used mice, we explicitly explained that any device was acceptable, and then removed the non-mouse users from the analysis.

Once workers accepted the task, they were asked to read the online instructions, which stated that they should perform the task as rapidly and accurately as possible. This was also always written at the top of the experimental window as a reminder (**Figure 2A**). After they finished all 25 sessions (including a practice session) and completed the questionnaire, the log data was exported to a csv file. They uploaded the file to a server and then received a payment of JPY 100 (~USD 0.92).

In total, 398 workers completed the task, including 384 mouse users according to the questionnaire results. Hereafter, we analyze only the mouse-users' data. The mouse users' demographics were as follows. Age: 16 to 76 years, with $M = 43.6$ and $SD = 11.0$. Gender: 300 male, 79 female, and 5 chose not to answer. Handedness: 24 were left-handed and 360 were right-handed. Windows version: 1 used Vista, 27 used Win7, 8 used Win8, and 348 used Win10. PC usage history: 0 (less than 1 year) to 45 years, with $M = 21.8$ and $SD = 7.82$.

In this study, we do not analyze these demographic data in detail. For example, it has been reported that participants' handedness (Hoffmann, 1997), gender and age (Brogmus, 1991) affect Fitts' law performance. In our simulation, it is possible that the data may be biased; e.g., when we select $N_P = 10$ workers, they are all males in their 60s. If researchers want to investigate this point, controlling the sampled workers' demographics before executing the simulation is needed.

For mouse users, the main pointing task took 3 min 45 s on average without breaks. With breaks, the mean task completion time was 5 min 42 s, and thus the effective hourly payment was JPY 1,053 (~USD 9.69). Note that this effective payment could change depending on other factors such as the times for reading the instructions and for uploading the csv file.

4. RESULTS

4.1. Outlier Data Screening

Following previous studies (MacKenzie and Isokoski, 2008; Findlater et al., 2017), we removed trial-level spatial outliers if the distance of the first click position was shorter than half of target distance $A/2$ (i.e., clicking closer to the non-target than the target) to omit clear accidental operations such as double-clicking the previous target. Another criterion used in these studies was to remove trials in which the click position was more than twice of target width $2W$ away from the target center. We did not use this criterion, as we would like to measure error trials even where a click position was $\geq (2W + 1)$ pixels away from the target center.

To detect trial-level temporal outliers to remove extremely fast or slow operations, we used the inter-quartile range (*IQR*) method (Devore, 2011), which is more robust than the *mean-and-3σ* approach. The *IQR* is defined as the difference between the third and first quartiles of the *MT* for each session for each participant. Trials in which the *MT* was more than $3 \times IQR$ higher than the third quartile or more than $3 \times IQR$ lower than the first quartile were removed.

For participant-level outliers, we calculated the mean *MT* across all 24 conditions ($3_A \times 8_W$) for each participant. Then, using each participant's mean *MT*, we again applied the *IQR* method and removed extremely rapid or slow participants. The trial- and participant-level outliers were independently detected and removed.

As a result, among the 92,160 trials ($= 3_A \times 8_W \times 10_{\text{repetitions}} \times 384_{\text{workers}}$), we identified 1,191 trial-level outliers (1.29%). We also found two participant-level outlier workers. While the mean *MT* of all participants was 898 ms and the *IQR* was 155 ms, the outlier workers' mean *MT*s were 1,462 and 1,533 ms. Accordingly, the data from all 480 trials of these two workers were removed ($= 3_A \times 8_W \times 10_{\text{repetitions}} \times 2_{\text{workers}}$). They also exhibited seven trial-level outliers (i.e., there were overlaps). In total, the data from 1,664 trials were removed (1.81%), which was close to the rate in a previous study (Findlater et al., 2017). As a result, we analyzed the remaining data from 90,496 trials.

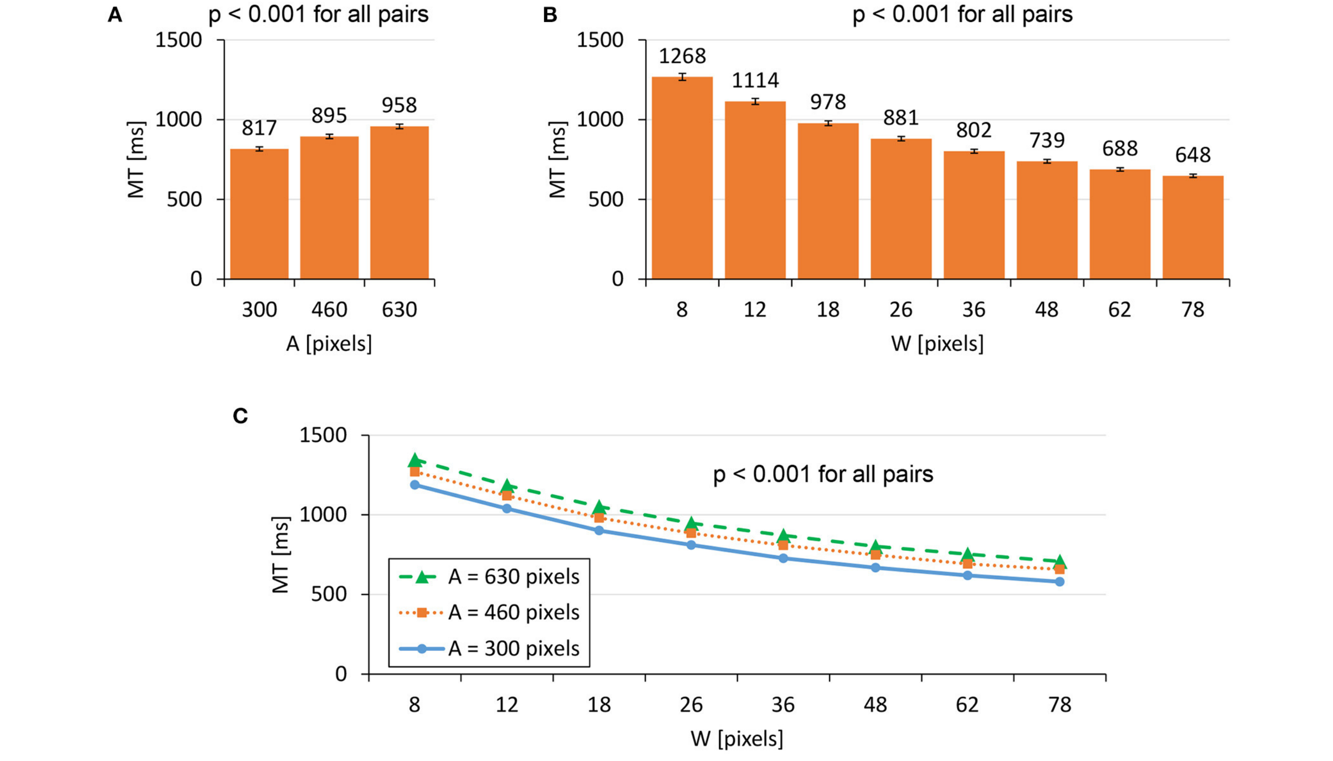


FIGURE 3 | Main effects of **(A)** target distance *A* and **(B)** target width *W* on *MT*. **(C)** The interaction effect of *A* × *W* on *MT*. Error bars indicate 95% confidence intervals.

4.2. Analyses of Dependent Variables

After the outliers were removed, the data from 90,496 trials (98.2%) were analyzed. The dependent variables were the MT, σ, and ER.

4.2.1. Movement Time

We used the Shapiro-Wilk test ($\alpha = 0.05$) and Q-Q plot to check the normality assumption required for parametric ANOVAs. The MT data did not pass the normality test, and thus we log-transformed the data to meet the normality assumption. The log-transformed data passed the normality test, and we used RM-ANOVAs with Bonferroni's p-value adjustment method for pairwise comparisons. For the F statistic, the degrees of freedom were corrected using the Greenhouse-Geisser method when Mauchly's sphericity assumption was violated ($\alpha = 0.05$).

We found significant main effects of A ($F_{1.909,727.1} = 2674$, $p < 0.001$, $\eta_p^2 = 0.88$) and W ($F_{4.185,1595} = 6813$, $p < 0.001$, $\eta_p^2 = 0.95$) on MT. A significant interaction was found for $A \times W$ ($F_{13.01,4955} = 14.23$, $p < 0.001$, $\eta_p^2 = 0.036$). **Figure 3** shows that the MT increased as A increased or W decreased. Regarding Fitts' law fitness, **Figure 4** shows that the model held well with $R^2 = 0.9789$. Previous studies using mice have reported that Fitts' law held with $R^2 > 0.9$ (Plamondon and Alimi, 1997; MacKenzie, 2013), and our dataset was consistent with these results.

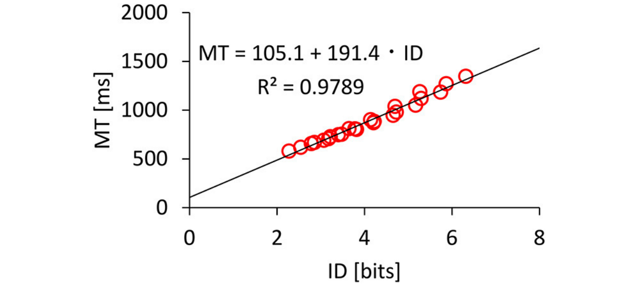


FIGURE 4 | Model fitness results for Fitts' law.

4.2.2. Click Point Variability

The σ data and its log-transformed data did not pass the normality test, and thus we used a non-parametric ANOVA with aligned rank transform (Wobbrock et al., 2011) with Tukey's p-value adjustment method for pairwise tests. We found significant main effects of A ($F_{2,762} = 3.683$, $p < 0.05$, $\eta_p^2 = 0.0096$) and W ($F_{7,2667} = 6043$, $p < 0.001$, $\eta_p^2 = 0.94$) on σ. An interaction of $A \times W$ was not significant ($F_{14,5334} = 0.8411$, $p = 0.62$, $\eta_p^2 = 0.0022$). **Figure 5** shows that the σ increased as A or W increased. The model fitness of Equation 2 ($\sigma = c + d \cdot W$) was quite high ($R^2 = 0.9966$), as shown in **Figure 6**. This fitness was greater than the results in previous studies, e.g., $R^2 = 0.9756$ (Bi and Zhai, 2013) and $R^2 = 0.9763$ (Yamanaka and Usuba, 2020)

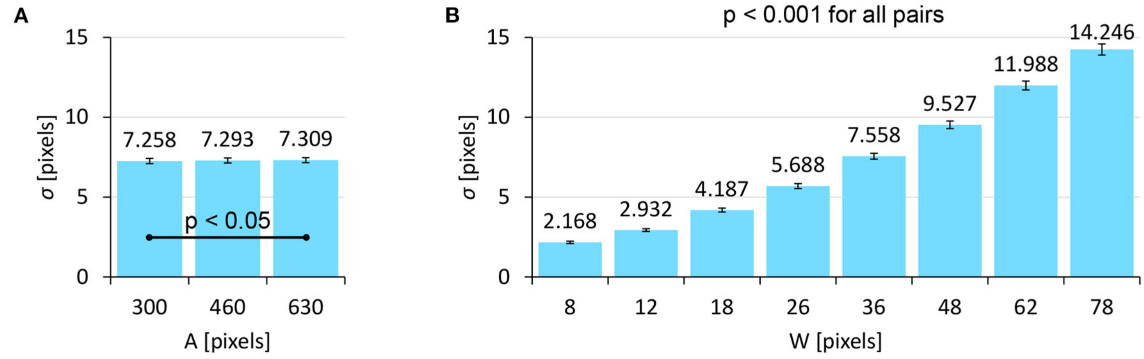


FIGURE 5 | Main effects of **(A)** target distance *A* and **(B)** target width *W* on σ. Error bars indicate 95% confidence intervals.

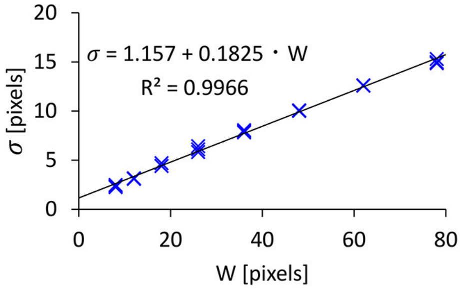


FIGURE 6 | Model fitness results for click point variability.

using touchscreens, and $R^2 = 0.9931$ using a virtual-reality input device (Oculus Touch wireless controller) (Yu et al., 2019).

Our model assumes that σ is not affected by A, but the result showed that A significantly affected σ. This statistical significance likely comes from the large number of participants. When we checked this in more detail, we found that the effect size of A was quite small compared with W ($\eta_p^2 = 0.0096$ vs. 0.94, respectively), and the mean σ values for $A = 300$, 460, and 630 pixels were 7.258, 7.293, and 7.309 pixels, which fall within a 0.051-pixel range (<1%). In contrast, the σ values varied from 2.168 to 14.25 pixels due to W (i.e., a 557% difference). While we plotted 24 points ($3_A \times 8_W$) in **Figure 6**, it looks as though there were only eight points, as the three σ values for the three As were almost the same and thus they overlapped.

4.2.3. Error Rate

The *ER* data and its log-transformed data did not pass the normality test, and thus we again used a non-parametric ANOVA with aligned rank transform. We found significant main effects of A ($F_{2,762} = 6.732, p < 0.01, \eta_p^2 = 0.017$) and W ($F_{7,2667} = 96.90$, $p < 0.001$, $\eta_p^2 = 0.20$) on *ER*. An interaction of $A \times W$ was not significant ($F_{14,5334} = 1.627, p = 0.064, \eta_p^2 = 0.0043$). **Figure 7** shows that the *ER* decreased as W increased, while A did not exhibit a clear tendency to increase/decrease the *ER*.

Using Equations 2 and 6, we can predict the *ER*s based on given W values. The predicted and actually observed *ER*s are shown in **Figure 8**. The worst prediction error was 4.235 points in the case of $(A, W) = (300, 8)$. As a comparison, previous studies on touch-based pointing tasks have reported that the prediction error for $W = 2.4$-mm targets was 9.74 points (Bi and Zhai, 2016) and that for 2-mm was 10.07 points (Yamanaka and Usuba, 2020). While a direct comparison with touch operations is not particularly fruitful, the tendency that prediction errors increase for smaller Ws is consistent between the previous studies and ours.

To formally evaluate our model's prediction accuracy, we computed the following three fitness criteria. The correlation between predicted vs. observed *ER*s was $R^2 = 0.9581$. The mean absolute error *MAE* was 1.193%. The root mean square error *RMSE* was 1.665%. In addition, to evaluate the prediction accuracy for new (unknown) task conditions, we ran a leave-one-(A, W)-out cross-validation. The three criteria for the *ER* prediction were $R^2 = 0.9529$, $MAE = 1.272\%$, and $RMSE = 1.814$. The worst prediction error was 4.805 points. These results indicate that, even for researchers who would like to predict the *ER* for a new task condition based on previously measured data, the prediction accuracy would not be considerably degraded.

5. SIMULATION

Although our N_{repeat} (10) was not large compared with previous studies on error-rate prediction models due to the time constraint for crowdsourcing, we hypothesized that increasing N_P would improve the model fitness. We also wonder how the model fitness changes when N_{repeat} is much smaller, which further shortens the task completion time for workers. For example, if it were 5, the average task completion time would be 2 min 51 s including breaks (i.e., half of 5 min 42 s). Note that N_{repeat} must be greater than 1 to compute the standard deviation σ.

We randomly selected N_P workers' data from the 384 mouse users by changing N_P from 10 (typical lab-based experiments) to 320 by doubling it repeatedly. The N_{repeat} changed from 2 to 10;

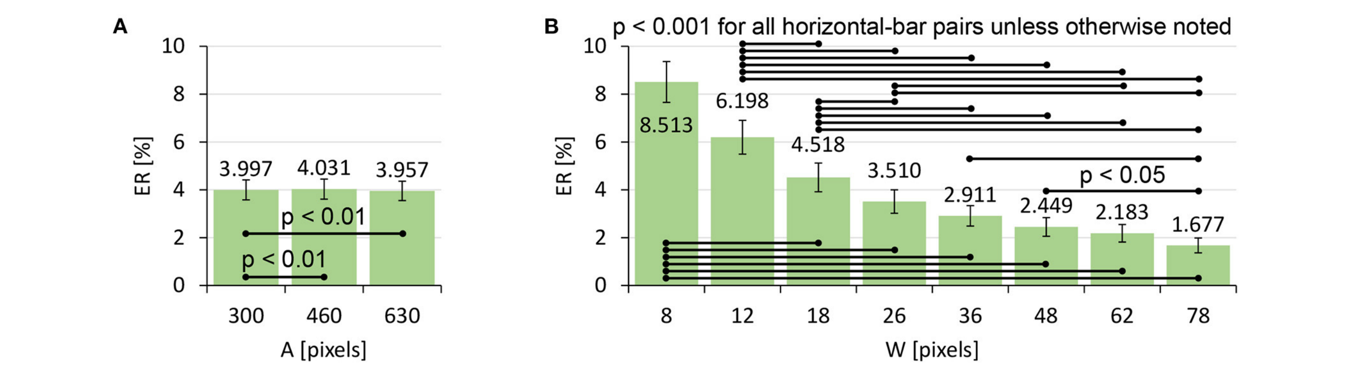


FIGURE 7 | Main effects of **(A)** target distance *A* and **(B)** target width *W* on *ER*. Error bars indicate 95% confidence intervals.

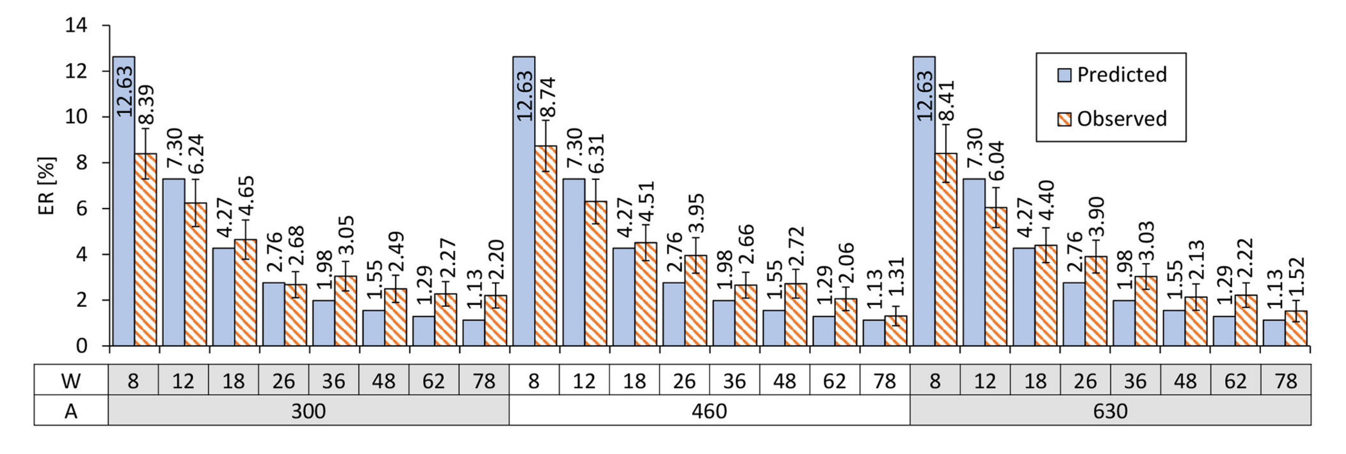


FIGURE 8 | Comparison of the predicted vs. observed *ER*s. Error bars indicate 95% confidence intervals.

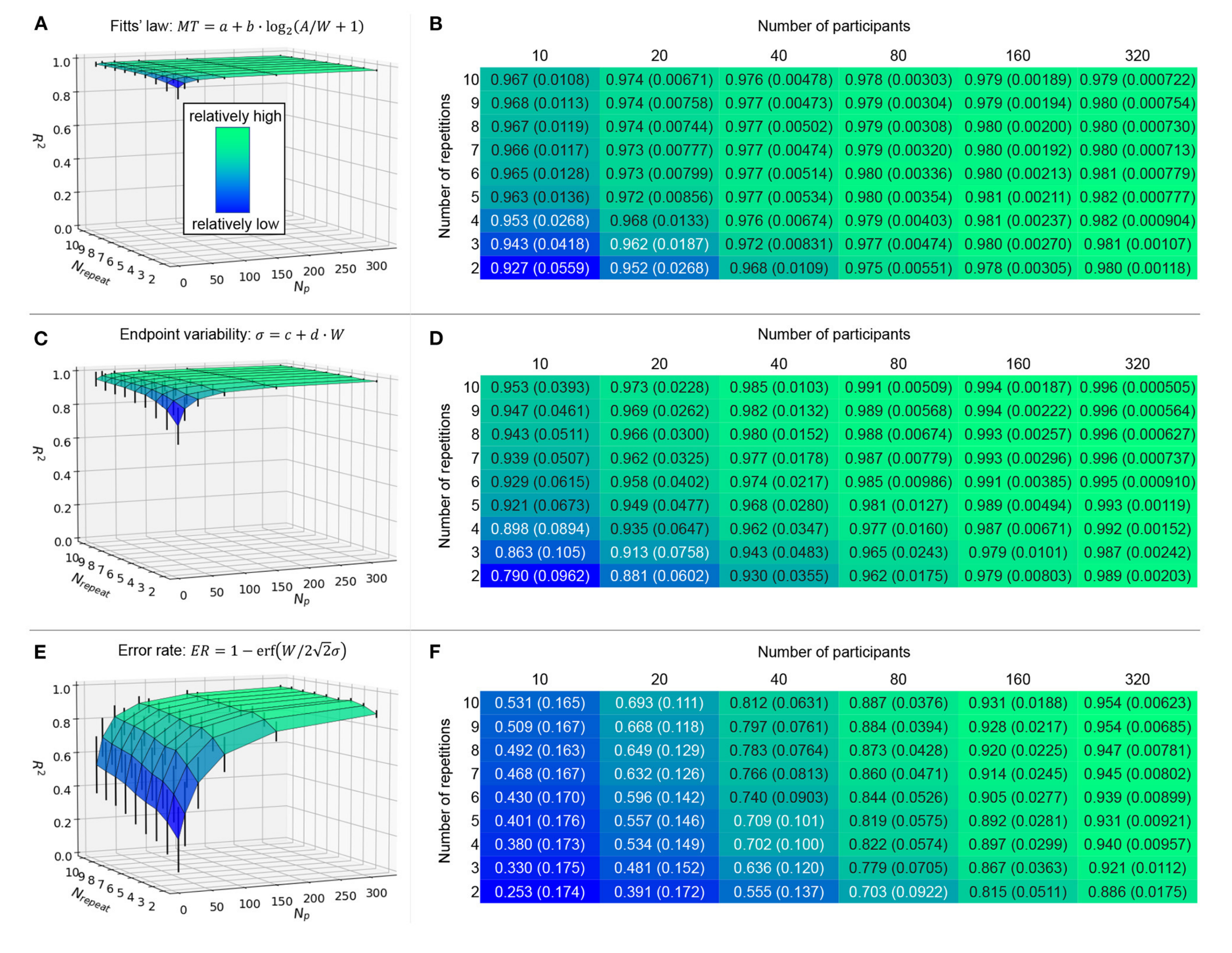
if it was 2, we used only the first two repetitions' data and the subsequent eight trials were removed. Outlier detection was run in the same manner as if we had conducted an experiment newly with N_P workers. Then, we analyzed the R^2 values for Equations 1 (Fitts' law), 2 (click point variability σ), and 6 (ER). To handle the randomness to select N_P workers, we ran this process over 1,000 iterations and computed the mean and *SD* values of the R^2s for each of $N_P \times N_{\text{repeat}}$.

The results are shown in **Figure 9**. First, we can visually confirm that the time prediction model (A) showed the flattest fitness compared with the other two models (C and E). The R^2 values were consistently over 0.92, and after we collected 20 participants or measured four repetitions, R^2 was over 0.95 (B). This result supports the decision of previous studies' lab-based experiments that recruited ten to 20 participants to examine Fitts' law. While repeating 15 to 25 trials per task condition has been recommended (Soukoreff and MacKenzie, 2004), our results show that a much smaller number of repetitions will suffice.

For the click point variability, as (C) shows, the model fitness was relatively worse only when both N_P and N_{repeat} are small. The increase in either N_P or N_{repeat} can resolve this. For example, by collecting $N_P \geq 80$ workers or repeating ten trials, we obtain $R^2 > 0.95$.

Lastly, for the error-rate model, the fitness was affected by N_P and N_{repeat} most drastically, as shown in (E). Particularly for small N_P values such as 10 and 20, the R^2 values were less than 0.70 (F), which is a unique result compared with the other two models that always showed much greater R^2 values in (B) and (D). If we fully use ten repetitions and would like to obtain a certain value of the model fitness (such as $R^2 > 0.9$), collecting 160 participants is sufficient—more precisely, when we tested N_P from 80 to 160 (step: 1), $N_P = 96$ achieved mean $R^2 = 0.9017 > 0.9$ for the first time ($SD = 0.03208$).

Figures 9E,F demonstrates that increasing N_P can be a viable alternative to increasing N_{repeat} to obtain a higher prediction accuracy for this error-rate model. Suppose we have a case where researchers want to set a smaller N_{repeat} such as 3 instead of 10 due to (e.g.,) asking workers to answer more questionnaire items after the task. Even for this case, by collecting $N_P = 320$ workers, the model would fit to the data with $R^2 > 0.9$ in our data. Hence, although the task completion time for crowdsourced user experiments should not be too long (Gould et al., 2016), the easy recruitment for crowdsourcing enables researchers to



B

Number of repetitions	Number of participants					
	10	20	40	80	160	320
10	0.967 (0.0108)	0.974 (0.00671)	0.976 (0.00478)	0.978 (0.00303)	0.979 (0.00189)	0.979 (0.000722)
9	0.968 (0.0113)	0.974 (0.00758)	0.977 (0.00473)	0.979 (0.00304)	0.979 (0.00194)	0.980 (0.000754)
8	0.967 (0.0119)	0.974 (0.00744)	0.977 (0.00502)	0.979 (0.00308)	0.980 (0.00200)	0.980 (0.000730)
7	0.966 (0.0117)	0.973 (0.00777)	0.977 (0.00474)	0.979 (0.00320)	0.980 (0.00192)	0.980 (0.000713)
6	0.965 (0.0128)	0.973 (0.00799)	0.977 (0.00514)	0.980 (0.00336)	0.980 (0.00213)	0.981 (0.000779)
5	0.963 (0.0136)	0.972 (0.00856)	0.977 (0.00534)	0.980 (0.00354)	0.981 (0.00211)	0.982 (0.000777)
4	0.953 (0.0268)	0.968 (0.0133)	0.976 (0.00674)	0.979 (0.00403)	0.981 (0.00237)	0.982 (0.000904)
3	0.943 (0.0418)	0.962 (0.0187)	0.972 (0.00831)	0.977 (0.00474)	0.980 (0.00270)	0.981 (0.00107)
2	0.927 (0.0559)	0.952 (0.0268)	0.968 (0.0109)	0.975 (0.00551)	0.978 (0.00305)	0.980 (0.00118)

D

Number of repetitions	Number of participants					
	10	20	40	80	160	320
10	0.953 (0.0393)	0.973 (0.0228)	0.985 (0.0103)	0.991 (0.00509)	0.994 (0.00187)	0.996 (0.000505)
9	0.947 (0.0461)	0.969 (0.0262)	0.982 (0.0132)	0.989 (0.00568)	0.994 (0.00222)	0.996 (0.000564)
8	0.943 (0.0511)	0.966 (0.0300)	0.980 (0.0152)	0.988 (0.00674)	0.993 (0.00257)	0.996 (0.000627)
7	0.939 (0.0507)	0.962 (0.0325)	0.977 (0.0178)	0.987 (0.00779)	0.993 (0.00296)	0.996 (0.000737)
6	0.929 (0.0615)	0.958 (0.0402)	0.974 (0.0217)	0.985 (0.00986)	0.991 (0.00385)	0.995 (0.000910)
5	0.921 (0.0673)	0.949 (0.0477)	0.968 (0.0280)	0.981 (0.0127)	0.989 (0.00494)	0.993 (0.00119)
4	0.898 (0.0894)	0.935 (0.0647)	0.962 (0.0347)	0.977 (0.0160)	0.987 (0.00671)	0.992 (0.00152)
3	0.863 (0.105)	0.913 (0.0758)	0.943 (0.0483)	0.965 (0.0243)	0.979 (0.0101)	0.987 (0.00242)
2	0.790 (0.0962)	0.881 (0.0602)	0.930 (0.0355)	0.962 (0.0175)	0.979 (0.00803)	0.989 (0.00203)

F

Number of repetitions	Number of participants					
	10	20	40	80	160	320
10	0.531 (0.165)	0.693 (0.111)	0.812 (0.0631)	0.887 (0.0376)	0.931 (0.0188)	0.954 (0.00623)
9	0.509 (0.167)	0.668 (0.118)	0.797 (0.0761)	0.884 (0.0394)	0.928 (0.0217)	0.954 (0.00685)
8	0.492 (0.163)	0.649 (0.129)	0.783 (0.0764)	0.873 (0.0428)	0.920 (0.0225)	0.947 (0.00781)
7	0.468 (0.167)	0.632 (0.126)	0.766 (0.0813)	0.860 (0.0471)	0.914 (0.0245)	0.945 (0.00802)
6	0.430 (0.170)	0.596 (0.142)	0.740 (0.0903)	0.844 (0.0526)	0.905 (0.0277)	0.939 (0.00899)
5	0.401 (0.176)	0.557 (0.146)	0.709 (0.101)	0.819 (0.0575)	0.892 (0.0281)	0.931 (0.00921)
4	0.380 (0.173)	0.534 (0.149)	0.702 (0.100)	0.822 (0.0574)	0.897 (0.0299)	0.940 (0.00957)
3	0.330 (0.175)	0.481 (0.152)	0.636 (0.120)	0.779 (0.0705)	0.867 (0.0363)	0.921 (0.0112)
2	0.253 (0.174)	0.391 (0.172)	0.555 (0.137)	0.703 (0.0922)	0.815 (0.0511)	0.886 (0.0175)

FIGURE 9 | Simulation results on mean (and *SD*) model fitness in R^2 by changing N_P and N_{repeat} over 1,000 iterations. Error bars indicate 1*SD*.

measure the central tendency of error rates. This benefit of crowdsourcing is more critical for error-rate models than time-prediction models, as we demonstrated here, which has never been empirically reported before.

When N_P or N_{repeat} was large, the error bars for model fitness (the *SD* values of R^2 over 1,000 iterations) were small for all models we examined (see **Figure 9**). This is because the same workers' data were more likely to be selected as the number of measured data points increased, and thus the variability in model fitness became small. In other words, when the number of data points was small, the model fitness depended more strongly on the choice of worker group and their limited trials. This effect of small N_P or N_{repeat} values on the large fitness variability was more clearly observed for the *ER* model (**Figures 9E,F**). Therefore, it is possible that the *ER* model will exhibit a quite low R^2 value when N_P or N_{repeat} was small, and at the same time, a much higher R^2 value might also be found by chance. This result shows that the *ER* is relatively not robust against the small number of data points.

In comparison, even when N_P or N_{repeat} was small, the error bars of the *MT* and σ models were smaller (**Figures 9A,C**). In particular, because the mean R^2 values of the *MT* model were already high (>0.92), there remains a limited space to exhibit much lower or higher R^2s, and thus the *SD* values could not be large. This demonstrated the robustness of the operational time prediction using Fitts' law.

6. DISCUSSION

6.1. Benefits and Implications of Using Crowdsourcing for Error-Rate Model Evaluation

In this study, we explored the potential of crowdsourcing for evaluating error-rate prediction models on GUIs. As one of the most fundamental operations, we utilized a Fitts' law task for its well-structured methodology. The results obtained from 384 crowdworkers showed that the models on Fitts' law and the click point variability fit well to the empirical data with $R^2 = 0.9789$ and 0.9966, respectively, as shown in **Figures 4**, **6**. Using the predicted σ values based on W, we then predicted the *ER*s for each $A \times W$ condition, which yielded the correlation between predicted vs. observed *ER*s of $R^2 = 0.9572$. The other metrics (*MAE* and *RMSE*) and the cross-validation also showed the good prediction accuracy of the model. On the basis of these results, in addition to the time-prediction model, we empirically demonstrated the first evidence that an error-rate model held well even for crowdsourced user experiments, even though it has been cautioned that crowdworkers are more error-prone in GUI tasks (Komarov et al., 2013; Findlater et al., 2017).

The simulation to alter N_P and N_{repeat} showed that the prediction accuracy of the error-rate model became better when either of these values was larger. This effect was more clearly observed for the error-rate model than the time- and click-point-variability models. In particular for the time model, the prediction accuracy reached close to the upper limit ($R^2 = 1$) even when the N_P and N_{repeat} were not large, such as the $R^2 > 0.95$ exhibited by ten workers performing four repetitions (**Figure 9B**). This suggests that the advantage of crowdsourcing in terms of its easy recruitment of numerous workers is not so critical. In comparison, for the error-rate model, increasing the N_P was still effective for $N_P \geq 160$.

Because the error rate is computed on the basis of occasionally occurring operations (clicking outside the target), researchers need more data to measure the theoretical value. Thus, our result, i.e., that collecting more data would lead to the theoretical value that a model estimates, is intuitive, but it has never been empirically demonstrated until now. Finally, our research hypothesis, "instead of increasing the number of repetitions per task condition, recruiting more workers is another approach to measure the error rates precisely, which will lead to a good prediction accuracy by the error-rate model," was supported. This is a motivating finding for future studies on evaluating novel error-rate models through crowdsourced user experiments.

Note that, we compared the sensitivity of time and error-rate models against N_P and N_{repeat}, but our purpose here was not to claim that (e.g.,) Fitts' law is a better model than the error-rate model. As described in the introduction, an *MT* is measured in every trial and then averaged after completing a session consisting of N_{repeat} trials, but an *ER* is computed after each session. Due to this difference, surmising that *the error-rate model is inferior* is not appropriate. Although more participants are needed to obtain a good fitness comparable with Fitts' law, which could be a limitation of the error-rate model, it does not necessarily mean that the model is wrong or inaccurate. Collecting numerous participants can avoid reaching such a mistaken conclusion. This point about making a conclusion based on an experiment with small sample size has been made before (Kaptein and Robertson, 2012; Caine, 2016), and our results again support the importance of a large sample size. Using crowdsourcing for error-rate model evaluation is a straightforward way to enable the recruitment of hundreds of participants with a reasonable time period, cost, and effort by researchers, which enhances the contribution of crowdsourcing to an undeveloped use application.

6.2. Limitations and Future Work

Our claims are limited to the task we chose and its design. We emphasized the usefulness of crowdsourced user experiments for error-rate model evaluation, but we only tested a GUI-task model implemented with mice following the Fitts' law paradigm. Within this scope, we limited the task design to horizontal movements where the effect of target height was negligible. We assume that modified models can predict *ER*s for more realistic targets such as pointing to circular targets (Bi and Zhai, 2016; Yamanaka and Usuba, 2020), but this needs further investigation in the future.

The model we examined was for selecting static targets, while recently models for more complicated tasks have been proposed, including those for pointing to automatically moving targets (Lee et al., 2018; Park and Lee, 2018; Huang et al., 2019), temporally constrained pointing such as rhythm games (Lee and Oulasvirta, 2016; Lee et al., 2018), and tracking a moving target (Yamanaka et al., 2020). We assume that the benefit of using crowdsourcing services to recruit numerous participants can be observed in these complicated tasks more clearly than our 1D pointing task. For example, pointing to a circular moving target needs more task parameters, such as the initial target distance A, its size W, movement speed V, and movement angle θ (Hajri et al., 2011; Huang et al., 2019). Because there are more task-condition combinations than 1D-target pointing, it is difficult to ask the participants to perform many repetitions per task condition, while recruiting numerous workers is easy in crowdsourced user studies. Investigating error rates in text input tasks is another important topic in the HCI field (Banovic et al., 2019; Cui et al., 2020) and would be a potential objective for crowdsourced user experiments.

A technical limitation specifically for our GUI-based experiment was that we could not check if workers really followed the given instruction, such as using mice and operating as rapidly and accurately as possible. For example, we fully trust the questionnaire results on the workers' devices. However, some mouse-users might use touchpads in actuality, as we had instructed to use mice. Similar concerns have been reported before: for touch pointing tasks with smartphones, researchers could not confirm whether workers tapped a target with their thumb as instructed (Yamanaka et al., 2019). Some other crowdsourcing platforms support an option that task requesters can ask workers to shoot a video when they perform a task, e.g., *UIScope* (http://uiscope.com/en). Still, this would create heavier workloads for both the workers and the experimenters. While these issues could not be completely removed at this time, if they were resolved in the future, the contribution to HCI would be significant.

7. CONCLUSION

We ran a crowdsourced user experiment to examine the benefits of recruiting numerous participants for evaluating an error-rate prediction model in a target pointing task, which is one of the most fundamental operations in PC usage. By analyzing the data obtained from 384 workers, we found that our model held well with $R^2 > 0.95$. Cross-validation also supported the good prediction accuracy to the unknown task conditions. In addition, when we randomly selected a limited portion of the entire workers from $N_P = 10$ to 320 and used only a limited number of trial repetitions from $N_{\text{repeat}} = 2$ to 10, we found that the time prediction model (Fitts' law) reached $R^2 > 0.95$ even if both of these values were small, while the error-rate model showed quite low fitness in that case. Thus, we empirically demonstrated that using crowdsourcing services for recruiting many participants is more clearly beneficial for evaluating the error-rate prediction model. Our findings should enhance the contribution of crowdsourcing in the HCI field.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the dataset used in this article is allowed to be open only in its statistically analyzed state (e.g., mean and standard deviation), and thus the raw dataset is not publicly available. Requests to access the datasets should be directed to SY, syamanak@yahoo-corp.jp.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Yahoo JAPAN Research's IRB-equivalent research ethics team. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SY has done all tasks required for preparing this article, including software development, data analyses, figure creation, and writing manuscript.

ACKNOWLEDGMENTS

The author would like to thank the anonymous reviewers of the AAAI HCOMP 2021 conference.

REFERENCES

Bailly, G., Lecolinet, E., and Nigay, L. (2016). Visual menu techniques. *ACM Comput. Surv.* 49, 1–41. doi: 10.1145/3002171

Banovic, N., Sethapakdi, T., Hari, Y., Dey, A. K., and Mankoff, J. (2019). "The limits of expert text entry speed on mobile keyboards with autocorrect," in *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '19* (New York, NY: Association for Computing Machinery), 1–12.

Batmaz, A. U., and Stuerzlinger, W. (2021). "The effect of pitch in auditory error feedback for fitts' tasks in virtual reality training systems," in *Conference on Virtual Reality and 3D User Interfaces, VR'21* (Lisbon), 1–10.

Bi, X., and Zhai, S. (2013). Bayesian touch: a statistical criterion of target selection with finger touch. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '13)*, 51–60.

Bi, X., and Zhai, S. (2016). "Predicting finger-touch accuracy based on the dual gaussian distribution model," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology UIST '16* (New York, NY: ACM), 313–319.

Brogmus, G. E. (1991). Effects of age and sex on speed and accuracy of hand movements: and the refinements they suggest for fitts' law. *Proc. Hum. Factors Soc. Annu. Meeting* 35, 208–212. doi: 10.1177/154193129103500311

Caine, K. (2016). "Local standards for sample size at chi," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '16* (New York, NY: Association for Computing Machinery), 981–992.

Casiez, G., and Roussel, N. (2011). "No more bricolage!: methods and tools to characterize, replicate and compare pointing transfer functions," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology UIST '11* (New York, NY: ACM), 603–614.

Cockburn, A., Lewis, B., Quinn, P., and Gutwin, C. (2020). "Framing effects influence interface feature decisions," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems CHI '20* (New York, NY: Association for Computing Machinery), 1–11.

Crossman, E. R. F. W. (1956). *The Speed and Accuracy of Simple Hand Movements*. Ph.D. thesis, University of Birmingham, Birmingham.

Cui, W., Zhu, S., Zhang, M. R., Schwartz, H. A., Wobbrock, J. O., and Bi, X. (2020). "Justcorrect: intelligent post hoc text correction techniques on smartphones," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology UIST '20* (New York, NY: Association for Computing Machinery), 487–499.

Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences, 8th Edn.* Boston, MA: Brooks and Cole publishing. Available online at: https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf

Do, S., Chang, M., and Lee, B. (2021). "A simulation model of intermittently controlled point-and-click behaviour," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems CHI '21* (New York, NY: Association for Computing Machinery), 1–17.

Findlater, L., Zhang, J., Froehlich, J. E., and Moffatt, K. (2017). "Differences in crowdsourced vs. lab-based mobile and desktop input performance data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems CHI '17* (New York, NY: ACM), 6813–6824.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *J. Exp. Psychol.* 47, 381–391.

Gould, S. J. J., Cox, A. L., and Brumby, D. P. (2016). Diminished control in crowdsourcing: an investigation of crowdworker multitasking behavior. *ACM Trans. Comput. Hum. Interact.* 23, 1–29. doi: 10.1145/2928269

Grossman, T., and Balakrishnan, R. (2005). "The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '05* (New York, NY: Association for Computing Machinery), 281–290.

Hajri, A. A., Fels, S., Miller, G., and Ilich, M. (2011). "Moving target selection in 2d graphical user interfaces," in *Human-Computer Interaction – INTERACT 2011*, eds P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler (Berlin: Springer), 141–161.

Hoffmann, E. R. (1997). Movement time of right- and left-handers using their preferred and non-preferred hands. *Int. J. Ind. Ergon.* 19, 49–57.

Huang, J., Tian, F., Fan, X., Tu, H., Zhang, H., Peng, X., and Wang, H. (2020). "Modeling the endpoint uncertainty in crossing-based moving target selection," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20* (New York, NY: Association for Computing Machinery), 1–12.

Huang, J., Tian, F., Fan, X., Zhang, X. L., and Zhai, S. (2018). "Understanding the uncertainty in 1d unidirectional moving target selection," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems CHI '18* (New York, NY: Association for Computing Machinery), 1–12.

Huang, J., Tian, F., Li, N., and Fan, X. (2019). "Modeling the uncertainty in 2d moving target selection," in *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology UIST '19* (New York, NY: Association for Computing Machinery), 1031–1043.

Jokinen, J. P. P., Sarcar, S., Oulasvirta, A., Silpasuwanchai, C., Wang, Z., and Ren, X. (2017). "Modelling learning of new keyboard layouts," in *Proceedings of the*

2017 CHI Conference on Human Factors in Computing Systems (New York, NY: Association for Computing Machinery), 4203–4215.

Kaptein, M., and Robertson, J. (2012). "Rethinking statistical analysis methods for chi," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '12* (New York, NY: Association for Computing Machinery), 1105–1114.

Komarov, S., Reinecke, K., and Gajos, K. Z. (2013). "Crowdsourcing performance evaluations of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13* (New York, NY: ACM), 207–216.

Lee, B., Kim, S., Oulasvirta, A., Lee, J.-I., and Park, E. (2018). "Moving target selection: A cue integration model," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18* (New York, NY: ACM), 230:1–230:12.

Lee, B., and Oulasvirta, A. (2016). "Modelling error rates in temporal pointing," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16* (New York, NY: ACM), 1857–1868.

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Hum. Comput. Interact.* 7, 91–139.

MacKenzie, I. S. (2002). *Bibliography of Fitts' Law Research.* Available online at: http://www.yorku.ca/mack/RN-Fitts_bib.htm(accessed August 24, 2021).

MacKenzie, I. S. (2013). A note on the validity of the shannon formulation for fitts' index of difficulty. *Open J. Appl. Sci.* 3, 360–368. doi: 10.4236/ojapps.2013.36046

MacKenzie, I. S. and Isokoski, P. (2008). "Fitts' throughput and the speed-accuracy tradeoff," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '08* (New York, NY: ACM), 1633–1636.

MacKenzie, I. S., Kauppinen, T., and Silfverberg, M. (2001). "Accuracy measures for evaluating computer pointing devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '01* (New York, NY: ACM), 9–16.

Matejka, J., Glueck, M., Grossman, T., and Fitzmaurice, G. (2016). "The effect of visual appearance on the performance of continuous sliders and visual analogue scales," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16* (New York, NY: Association for Computing Machinery), 5421–5432.

Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., and Smith, J. E. K. (1988). Optimality in human motor performance: ideal control of rapid aimed movements. *Psychol. Rev.* 95, 340–370.

Park, E., and Lee, B. (2018). Predicting error rates in pointing regardless of target motion. *arXiv [Preprint]*. arXiv: 1806.02973. Available online at: https://arxiv.org/pdf/1806.02973.pdf (accessed April 24, 2020).

Plamondon, R., and Alimi, A. M. (1997). Speed/accuracy trade-offs in target-directed movements. *Behav. Brain Sci.* 20, 279–303.

Schwab, M., Hao, S., Vitek, O., Tompkin, J., Huang, J., and Borkin, M. A. (2019). "Evaluating pan and zoom timelines and sliders," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY: Association for Computing Machinery), 1–12.

Soukoreff, R. W., and MacKenzie, I. S. (2004). Towards a standard for pointing device evaluation, perspectives on 27 years of fitts' law research in hci. *Int. J. Hum. Comput. Stud.* 61, 751–789. doi: 10.1016/j.ijhcs.2004.09.001

Wobbrock, J. O., Cutrell, E., Harada, S., and MacKenzie, I. S. (2008). "An error model for pointing based on fitts' law," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08* (New York, NY: ACM), 1613–1622.

Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (2011). "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,CHI '11* (New York, NY: ACM), 143–146.

Yamanaka, S. (2021a). "Comparing performance models for bivariate pointing through a crowdsourced experiment," in *Human-Computer Interaction – INTERACT 2021* (Gewerbestr: Springer International Publishing), 76–92.

Yamanaka, S. (2021b). "Utility of crowdsourced user experiments for measuring the central tendency of user performance to evaluate error-rate models on guis," in *AAAI HCOMP 2021* (Palo Alto, CA: AAAI), 1–12.

Yamanaka, S., Shimono, H., and Miyashita, H. (2019). "Towards more practical spacing for smartphone touch gui objects accompanied by distractors," in *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces, ISS '19* (New York, NY: Association for Computing Machinery), 157–169.

Yamanaka, S., and Usuba, H. (2020). Rethinking the dual gaussian distribution model for predicting touch accuracy in on-screen-start pointing tasks. *Proc. ACM Hum. Comput. Interact.* 4, 1–20. doi: 10.1145/3427333

Yamanaka, S., Usuba, H., Takahashi, H., and Miyashita, H. (2020). "Servo-gaussian model to predict success rates in manual tracking: Path steering and pursuit of 1d moving target," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, UIST '20* (New York, NY: Association for Computing Machinery), 844–857.

Yu, D., Liang, H.-N., Lu, X., Fan, K., and Ens, B. (2019). Modeling endpoint distribution of pointing selection tasks in virtual reality environments. *ACM Trans. Graph.* 38, 1–13. doi: 10.1145/3355089.3356544

Zhai, S., Kong, J., and Ren, X. (2004). Speed-accuracy tradeoff in fitts' law tasks: on the equivalency of actual and nominal pointing precision. *Int. J. Hum. Comput.Stud.* 61, 823–856. doi: 10.1016/j.ijhcs.2004.09.007

Zhao, J., Soukoreff, R. W., Ren, X., and Balakrishnan, R. (2014). A model of scrolling on touch-sensitive displays. *Int. J. Hum. Comput. Stud.* 72, 805–821. doi: 10.1016/j.ijhcs.2014.07.003

Conflict of Interest: SY is employed by Yahoo Japan Corporation.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 01 April 2022
doi: 10.3389/frai.2022.818451

Scaling and Disagreements: Bias, Noise, and Ambiguity

Alexandra Uma[1]*, *Dina Almanea*[1] and *Massimo Poesio*[1,2,3]

[1] Computational Linguistics Lab, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom, [2] Digital Environment Research Institute, Queen Mary University of London, London, United Kingdom, [3] Turing Institute, London, United Kingdom

Crowdsourced data are often rife with disagreement, either because of genuine item ambiguity, overlapping labels, subjectivity, or annotator error. Hence, a variety of methods have been developed for learning from data containing disagreement. One of the observations emerging from this work is that different methods appear to work best depending on characteristics of the dataset such as the level of noise. In this paper, we investigate the use of an approach developed to estimate noise, temperature scaling, in learning from data containing disagreements. We find that temperature scaling works with data in which the disagreements are the result of label overlap, but not with data in which the disagreements are due to annotator bias, as in, e.g., subjective tasks such as labeling an item as offensive or not. We also find that disagreements due to ambiguity do not fit perfectly either category.

Keywords: overlapping labels, annotation disagreement, observer disagreement, temperature scaling, model calibration, cost-sensitive loss

OPEN ACCESS

Edited by:
Matt Lease,
University of Texas at Austin,
United States

Reviewed by:
Christopher Welty,
Google, United States
Alexander Braylan,
University of Texas at Austin,
United States

****Correspondence:***
Alexandra Uma
alexandra.uma2@gmail.com

Specialty section:
This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

Received: *19 November 2021*
Accepted: *01 March 2022*
Published: *01 April 2022*

Citation:
Uma A, Almanea D and Poesio M (2022) Scaling and Disagreements: Bias, Noise, and Ambiguity. Front. Artif. Intell. 5:818451. doi: 10.3389/frai.2022.818451

1. INTRODUCTION

Crowdsourced data are often rife with disagreements between coders. Hence, a variety of methods have been developed for learning from data containing disagreement. In a previous study, the focus was on developing methods for removing items on which annotators disagreed (Beigman-Klebanov and Beigman, 2009), or aggregation methods able to learn "ground truth" from such data (Dawid and Skene, 1979; Smyth et al., 1994; Carpenter, 2008; Whitehill et al., 2009; Hovy et al., 2013) (see, e.g., Sheshadri and Lease, 2013; Paun et al., 2018, 2022; Uma et al., 2021b for review). More recent work however suggests that better results are obtained by methods training directly from data containing disagreements (Raykar et al., 2010; Rodrigues and Pereira, 2017; Peterson et al., 2019; Uma et al., 2020; Fornaciari et al., 2021; Uma et al., 2021b). But another finding emerging from this recent work is that different methods for learning from data containing disagreements work best depending on the dataset (Uma et al., 2021b). One possible explanation for this difference in performance is disagreements can be due to a number of causes, ranging from annotator error to problematic annotation schemes (e.g., with overlapping labels) to genuine item ambiguity to more general item difficulty. An early proposal regarding distinguishing between different types of disagreement was made by Reidsma and Carletta (2008), who showed that disagreements due to **(random) noise**—random annotator errors—affect model training differently from disagreements due to **bias**—annotator-dependent patterns. Such work raises the question of whether it is possible to distinguish between these two types of disagreement (or other types perhaps) so as to decide which method for learning from disagreement is more appropriate for a given dataset.

In early work (Uma et al., 2021b), we considered a number of approaches to identify the type of disagreement that was most typical in a dataset. However, the objective of the measures used in that work is to identify the type of disagreement in a dataset prior to training a model. In this paper, we report on an investigation of the use of an approach inspired by the idea of **temperature scaling** developed by, e.g., Platt (1999) and Guo et al. (2017) to allow a model to *automatically* adapt in the presence of disagreement in the data. We use a range of datasets known to contain disagreements arising from different sources (Uma et al., 2021b) to train models using the state-of-the-art **soft-loss** approach for learning from disagreement (Peterson et al., 2019; Uma et al., 2020, 2021b) and test whether adding automatic temperature scaling improves model performance. We find that the datasets used can be divided into three groups on the basis of the results obtained with the proposed approach. Automatic temperature scaling works well with datasets in which disagreement is mostly due to substantial overlap between the labels such that annotators have to choose a label more or less randomly. By contrast, the approach does not work at all with data in which the disagreements are due to a clear bias, as in, e.g., subjective tasks such as labeling an item as offensive or not, which is known to be affected by the annotators' political views. Finally, with datasets where most or part of the disagreement arises from linguistic ambiguity lie in between these extremes, suggesting that ambiguity may not sit perfectly within a binary distinction such as the distinction between bias and noise proposed by Reidsma and Carletta (2008).

2. METHODOLOGY: TEMPERATURE-SCALED SOFT LOSS

In this section, we introduce the **temperature-scaled soft loss** approach, which combines the soft loss approach to learning from disagreement we developed in previous work with our own approach to adding temperature scaling in a deep learning model, which we call **automatic temperature scaling**. We first review the soft-loss approach proposed by Peterson et al. (2019) and Uma et al. (2020) and extend soft-loss by including exploration of the suitability of various standard loss functions for soft-loss training. Next, we discuss the (automatic) temperature-scaled soft-loss methodology which involves weighting the soft loss for each item by a learned temperature parameter.

2.1. Soft Loss Learning

The soft-loss functions approach to training from data containing disagreement combines using a standard loss function with a probabilistic soft label generated from crowd annotations (Peterson et al., 2019; Uma et al., 2020). To train a model using the soft-loss function approach, a standard loss function such as cross-entropy or squared error is used; but instead of targeting the ground truth viewed as a one-hot label, a **soft label**—a probability distribution over the labels—is generated from the distribution of crowd labels and used as a target for training the machine learning model. We discuss each step in turn.

2.1.1. Generating Probabilistic Soft Labels

While experimenting with a variety of datasets standardly used for learning from disagreement, Uma et al. (2020) showed that for a soft-loss function, the quality of the predictions is dependent on the method used in generating the probabilistic soft labels, which in turn is dependent on the characteristics of the annotation for the dataset. They evaluated two standard label generation functions—the softmax function and the standard normalization function—finding which is best depends on the dataset. Soft labels obtained through standard normalization were found to be preferable for datasets like CIFAR-10H (Peterson et al., 2019), which were annotated by a large number of expert annotations with high observed agreement among them. Soft labels produced using softmax proved instead more suitable for datasets that do not meet these criteria, such as Gimpel et al.'s POS dataset (Plank et al., 2014a) and the LABELME dataset (Rodrigues and Pereira, 2017). Uma et al. (2021b) further showed that the best soft label for mixed quality datasets, such as PDIS (Poesio et al., 2019),

were obtained by using the posterior distribution of a probabilistic aggregation model such as MACE (Hovy et al., 2013). For our novel misogyny dataset ArMIS (Almanea and Poesio, 2022), we found that the normalized distribution of the annotators was the best-performing label.

2.1.2. A Suitable Loss Function

Peterson et al. (2019) only used the cross-entropy loss function, hypothesizing that it was uniquely suitable for the task. Uma et al. (2021b) tested a variety of other loss functions, including Kullback-Leibler (henceforth: KL) and (Summed) Squared Error (henceforth: SE)[1]. Malinin and Gales (2019) argued that for datasets with high noise due to overlapping labels and resulting in a multi-modal label distribution[2] reverse KL-divergence is most appropriate if the goal is to maximize prediction accuracy. They tested their hypothesis on synthetic data, comparing reverse KL-divergence as a loss function with (forward) KL divergence, and showed that while KL-divergence is a sensible loss function for datasets with low data uncertainty and target distributions where "correct" labels are available, reverse KL-divergence is more suitable when this is not the case.

Thus, as a preliminary experiment, we tested the hypothesis of Malinin and Gales (2019) with our (non-artificial) data by training soft-loss functions for each task using the best soft label and each of the divergence functions. We additionally tested the other two well-known probability-comparing loss functions—the cross-entropy loss function (CE) already used in Peterson et al. (2019) and Uma et al. (2020, 2021b) and the Squared error function (SE) used in Uma et al. (2021b). Soft-loss functions using each of the stated functions can be expressed using the simplified notation:

[1] After some experimentation, and in keeping with the other loss function, we decided to use the sum of the squared errors as opposed to the mean.

[2] Malinin and Gales (2019) use the term **data uncertainty** for this type of noise, but as far as we know their notion of data uncertainty is the same as what Reidsma and Carletta (2008) call random noise.

- Cross-Entropy Soft loss:

$$CE(y_{hum},\ y_\theta) = -\sum_{i=1}^{n} y^i_{hum} \log y^i_\theta \tag{1}$$

- KL Soft loss:

$$D_{KL}(y_{hum} \mid\mid y_\theta) = -\sum_{i=1}^{n} y_{hum} \log(\frac{y^i_\theta}{y^i_{hum}}) \tag{2}$$

- Reverse KL Soft loss[3]:

$$D_{RKL}(y_\theta \mid\mid y_{hum}) = \sum_{i=1}^{n} y^i_\theta \log(\frac{y^i_{hum}}{y^i_\theta}) \tag{3}$$

- SE Soft loss:

$$MSE(y_{hum}, y_\theta) = \sum_{i=1}^{n} (y^i_{hum} - y^i_\theta)^2 \tag{4}$$

where y^i_{hum} is the target label for an item i, the *best soft label*; y^i_θ is the model's predicted probability distribution for that item; and n is the number of items in the training set.

We experiment with these variations of the soft loss function and note the prediction accuracy of the trained models, especially in reaction to Malinin and Gales's (2019) hypothesis. The best soft loss function is used for experiments in automatic temperature scaling.

2.2. Item Weighting Through Automatic Temperature Scaling

One of the most widely adopted approaches to learning from disagreement involves developing methods for identifying **difficult** items–items on which there is an unexpected degree of disagreement among annotators. Such methods typically use statistical inference to infer the difficulty of an item, and then use such difficulty to weigh or filter items classified as intrinsically difficult (refer to, e.g., Carpenter, 2008; Beigman and Beigman Klebanov, 2009; Whitehill et al., 2009) and the discussion of item difficulty approaches in Paun et al. (2022). In the deep learning literature, a number of methods of this type were developed, for which the term **temperature scaling** is often used.

In this paper, we introduce a method of this type, which we called **automatic temperature scaling**, and combine ideas from both temperature scaling and **Platt scaling**. Platt scaling was proposed as a way to calibrate a logistic regression model, i.e., adjust its parameters to reflect uncertainty (Platt, 1999). To calibrate a model, Platt proposes that two **scalar parameters**, a and b ∈ R, be learned by optimizing the negative log-likelihood function over the validation set while keeping the model's parameters fixed. The learned parameters are used to rescale the logits of the model, $\mathbf{z}_i$ resulting in outputs, $f(\mathbf{x}_i) = \sigma(a\mathbf{z}_i + b)$.

Temperature scaling is a single parameter variant of Platt scaling (Guo et al., 2017), where a single scalar parameter, T, called the **temperature**, is used to rescale logit scores for all the classes, $\mathbf{z}_i$, before applying the softmax function. This way, the model's recalibrated probabilities are given as:

$$f(\mathbf{x}_i) = \sigma(\mathbf{z}_i/T) \tag{5}$$

where $\sigma(\cdot)$ is the softmax function. When $T > 1$, the entropy of the output probabilities increases, hence "softening the softmax" and evening out the probability distribution. $T < 1$ hardens the softmax, resulting in a peakier (more modal) probability distribution. Finally, $T = 1$ recovers the unscaled probabilities (Guo et al., 2017). The value of T is obtained by minimizing the negative log-likelihood on a held-out validation dataset. Because T is independent of the class, j, and the item, i, *temperature scaling does not affect which class is predicted and hence does not affect prediction accuracy*.

Automatic temperature scaling, which we propose here, is a natural extension of temperature scaling. It differs from standard temperature scaling in three key ways. First, automatic temperature scaling learns a parameter *vector* T_i jointly as it learns to predict the classes. It does this by learning a network of weights $\mathbf{w}_{T_i}$ and biases b_{T_i} such that

$$T_i = softplus(\mathbf{W}_{T_i}\mathbf{x}_i + b_{T_i)} \tag{6}$$

This network of weights is disjoint from the network of weights for learning to map inputs to targets. By using Softplus as the squashing the function (as opposed to sigmoid, ReLu, or Tanh) we apply non-linearity to the network without overly limiting the bounds of T_i[4].

The reason for moving from a single scalar parameter to a vectorial parameter, and from a single value for the whole corpus to an item dependent parameter, is that difficulty is very much item dependent—e.g., not all images are equally easy or difficult—and also class dependent: some classes are more easily confused than others, as discussed in more detail in the next section. The vectorial expression of temperature is similar to the one used in **matrix scaling**, an alternative temperature scaling also proposed by Guo et al. (2017)[5]. But unlike in matrix scaling (or Platt scaling, in which more than one parameter is also learned), the parameters are not tuned on a held-out validation set; rather, the model jointly learns classifier and scaling parameters. During training, the model's outputs, $\hat{y}_i = f(x_i)$ are computed as follows:

$$f(\mathbf{x}_i) = \sigma(\mathbf{z}_i * T_i) \tag{7}$$

The model's loss is computed using the appropriate soft loss function.

The second key difference is practical in nature but has notable implications. Unlike in temperature scaling, where the logits are divided by temperature T, in automatic temperature scaling,

[3] In reverse KL, the target human-derived soft label and the predicted soft label are swapped.

[4] Sigmoid, ReLu, and Tanh outputs are bounded between [0, 1], [0,1], and [−1, 1] respectively, while Softplus outputs are only lower bounded are zero.

[5] Guo et al. (2017) propose the use of the $max(\cdot)$ function, rather than $softplus(\cdot)$.

the logits are *multiplied* by the temperature; we found this to work better in practice. The consequence is that in automatic temperature scaling, a warmer temperature (higher values of T_i) indicates *lower* uncertainty resulting in peakier probabilities, while colder temperatures indicate higher uncertainty resulting in a more even distribution–the opposite to temperature scaling[6].

The third key difference can be observed from the definition of T_i in Equation (6). Unlike in standard temperature scaling, in automatic temperature scaling, the model does not have a single temperature value; rather, the temperature of any given item is a function of the input vector for the item and the temperature weights of the model, W_{T_i}—the logits for each instance are scaled to a different temperature, determined by the model and learned as a function of the input features of the instance. In this way, if the model is able to identify uncertainty for an input item, it will respond by producing a lower temperature value for that item. The converse is also true. Thus, by considering each instance separately, the model is able to produce temperature values depending on how much data uncertainty it perceives for each item.

This third aspect is vital to understanding the anticipated improvement in predictive accuracy using automatic temperature scaling. In datasets with overlapping labels, because the modal class for affected items is arbitrary, models (much like annotators) are likely to disagree with the modal class of the target labels, predicting a different (and possibly equally plausible) modal class for perceived noisy inputs. The temperature lowering for such items results in a flatter predicted probability distribution and has the added effect of decreasing the loss contribution of that item to the overall loss. Consequently, the model penalizes itself less for such items and reduces the loss contribution of the item to the total loss. In this way, automatic temperature scaling can be comparable to cost-sensitive loss (Plank et al., 2014a).

3. THE EXPERIMENTS

In this section, we present our experimental design and discuss the datasets and models used for the experiments conducted in this study.

3.1. Experiment Design

We conducted the experiments in two phases. First, we experimentally compared the suitability of various standard loss functions for soft loss training as outlined in Section 2.1.2 on several tasks. Then, we extended the best-performing loss function into an automatic temperature-scaled soft loss. For both experiments, we evaluated the models using two evaluation metrics, one hard and one soft.

3.1.1. Hard Evaluation

As a hard evaluation metric, we used accuracy, as done by Peterson et al. (2019) and Uma et al. (2020). We calculated the accuracy of each model's prediction with respect to a standard: the majority vote aggregate of the expert annotators for ArMIS [7] and gold labels for the other datasets.

3.1.2. Soft Evaluation

As noted in previous work (Dumitrache et al., 2018; Peterson et al., 2019; Uma et al., 2020; Basile et al., 2021; Uma et al., 2021b), as the realization that gold labels are an idealization growth, so does the awareness that hard evaluation is not sufficient to compare machine learning models on tasks in which disagreements are extensive, and extremely questionable for tasks in which the labels are subjective and therefore it does not make sense a "gold label" exists that the disagreements can be reconciled to. A particularly obvious illustration of this last point is the **misogyny detection** task, related to hate speech detection. In this task, the labels assigned by annotators are very much dependent on their background, i.e., text found misogynistic by a female annotator or a more liberal annotator may not be found misogynistic by a male annotator or an annotator from a more conservative background.

When evaluating tasks containing disagreements, or in which disagreements may be intrinsic, it would seem insightful not to evaluate models against a questionable gold label only, but also against **soft labels** in the sense discussed above (probability distributions over the labels derived from crowd annotations) in which disagreements are preserved. Consequently, in this paper, our models are also evaluated using a soft evaluation metric, cross-entropy. Like Peterson et al. and Uma et al., we compute the cross-entropy between the probability distribution produced by each model and the **best soft label** produced from the crowd distribution (The label that is most appropriate for that dataset, as discussed above). This form of evaluation provides insight into how well the models are able to capture possible disagreements in labeling resulting from the crowd.

3.2. Data

We used in this study four disagreement-preserving datasets that have been previously used in research into learning to classify from disagreement (Jamison and Gurevych, 2015; Plank et al., 2014a,b; Uma et al., 2020, 2021a; Fornaciari et al., 2021) and that exemplify different sources of disagreement (An in-depth analysis of the disagreements in these datasets has been carried out by Uma et al., 2021b). In addition, we used an entirely new dataset, ArMIS (Almanea and Poesio, 2022), illustrating a different type of disagreement not considered by Uma et al. (2021b): disagreement due to subjectivity.

3.2.1. The Gimpel et al. pos Corpus

The first example of a corpus containing disagreements due to ambiguity (Plank et al., 2014b) is Gimpel et al.'s (2011) POS dataset (henceforth, POS), which has been often used in research into developing disagreement-aware NLP models (Plank et al., 2014a; Jamison and Gurevych, 2015; Fornaciari et al., 2021; Uma et al., 2021b). The dataset consists of 14k Twitter posts annotated with ground truth POS tags collected by Gimpel et al. (2011) from

[6] As such, it would be more appropriate to name T_i "confidence" or "certainty"—but we will stick with the original name to acknowledge the intellectual debt of our proposal to temperature scaling.

[7] Ties were broken by making a random selection.

expert annotators and crowdsourced tags collected by Plank et al. (2014b)—at least five crowdsourced labels per token from 177 annotators.

The workers annotating this corpus often disagree with the ground truth label; the observed agreement (A_o, Artstein and Poesio, 2008) for the dataset is 0.73, as computed using the multi-annotator version of Fleiss Kappa (Fleiss et al., 2004).

A typical example of the disagreements found in this corpus is shown below (the token to be tagged is in bold):

(8)	Noam	likes	**social**	media
	Noun	Verb	Adj/Noun	Noun

in the context, the category *Noun* would seem to be just as appropriate as the category *Adj* for the token ***social***.

Plank et al. (2014b) conducted an analysis of the easy and hard cases in this dataset, finding that the vast majority of inter-annotator disagreements are due to **genuine linguistic ambiguity**, as in this example, although the POS categories Adj and Noun are clearly distinct, in some cases, it is not possible to tell what is the "right" category (Plank et al., 2014b). In fact, an analysis of the POS dataset carried out by Uma et al. (2021b) showed that the average observed agreement on an "easy" category such as nouns (particularly for name tokens like Twitter handles) is much higher than for other categories.

For experiments using this dataset, we split the 14k tokens into training (12k) and testing (2k) and use the development dataset released by Plank et al. (2014a) for validation.

3.2.2. The PDIS Corpus

The second corpus we used contains disagreements in part due to ambiguity, in part to annotator carelessness. The *Phrase Detectives* 2 corpus (Poesio et al., 2019) is a crowdsourced anaphoric reference corpus collected with the *Phrase Detectives* game-with-a-purpose (Poesio et al., 2013)[8]. Anaphoric reference is another aspect of linguistic interpretation in which ambiguity is rife (Poesio et al., 2006; Versley, 2008; Recasens et al., 2011). For example, Poesio et al. (2006) discussed examples such as (3.2.2).

(9)
```
3.1     M: can we .. kindly hook up
3.2      : uh
3.3      : engine E2 to the boxcar at ..
           Elmira
4.1     S: ok
5.1     M: +and+ send \textcolor{red}
           {\textbf{it}} to Corning
5.2      : as soon as possible please
6.1     S: okay
         [2sec]
7.1     M: do let me know when it gets
           there
8.1     S: okay it'll /
8.2      : it should get there at 2 AM
9.1     M: great
9.2      : uh can you give the
9.3      : manager at Corning instructions
           that
9.4      : as soon as it arrives
9.5      : it should be filled with
           oranges
10.1    S: okay
10.2     : then we can get that filled
```

In this exchange, it is not clear whether the pronoun *it* in 5.1 (in red) refers to *the engine E2* that has been hooked up to *the boxcar at Elmira* or to the boxcar itself or indeed whether the distinction matters at all. It is only at utterance 9.5 that we get evidence that *it* probably refers to *the boxcar at Elmira* since only boxcars can be filled with oranges. The two interpretations are clearly distinct–the pronoun cannot refer to both–but it is not possible to decide which is the intended one from the context.

The *Phrase Detectives* 2 corpus consists of 542 documents, for a total of 408K tokens and 107K markables, annotated by slightly less than 2,000 players producing a total of 2.2M judgments—about 20 judgments per markable on average. In total, 64.3% of the markables received more than one distinct interpretation from the players. Some of the disagreements are due to annotator error/carelessness, others to interface issues; but for about 10% of markables, disagreement is again due to **genuine linguistic ambiguity**.

In this study, we used PDIS, a simplified version of the corpus containing only binary information status labels: discourse new (DN) (the entity referred to has never been mentioned before) and discourse old (DO) (it has been mentioned). PDIS still consists of 542 documents, for a total of 408K tokens and over 96K markables; an average of 11.87 annotations per markable are preserved[9].

Forty-five of the documents (5.2K markables), collectively called PD_{gold}, additionally contain expert-adjudicated gold labels. This subset of PDIS was designated as the test set. The training and development datasets consist of 473 documents (and 86.9K markables) and 24 documents (4.2K markables), respectively[10].

3.2.3. The LabelMe Corpus

The most widely used corpus for learning to classify images from crowds is the LabelMe dataset[11] (Russell et al., 2008). It classifies outdoor images according to 8 categories: *highway, inside city, tall building, street, forest, coast, mountain*, or *open country*. Using Amazon Mechanical Turk, Rodrigues and Pereira (2017) collected an average of 2.5 annotations per image from 59 annotators for 10K images in this dataset.

The observed agreement for this dataset, also computed using the multi-annotator version of Fleiss et al.'s (2004) Kappa, is 0.73, which is the same level of average observed agreement seen in the POS dataset. However, it can be argued that the source and nature of the disagreement in this dataset are different, consider **Figure 1** for an illustration. The ground truth label for the example image is *inside city*, and one annotator chose that label as well, but two other annotators chose *tall building*. Notice the difference from the ambiguity cases in POS and PDIS: there, two interpretations are possible, but a word can only have one—it is just that it is

[8] https://github.com/dali-ambiguity

[9] DO judgments with different antecedents are considered identical, and the judgments other than DN or DO are removed.

[10] Another example of corpus were the disagreement is due to linguistic ambiguity is Dumitrache et al. (2019).

[11] http://labelme.csail.mit.edu/Release3.0

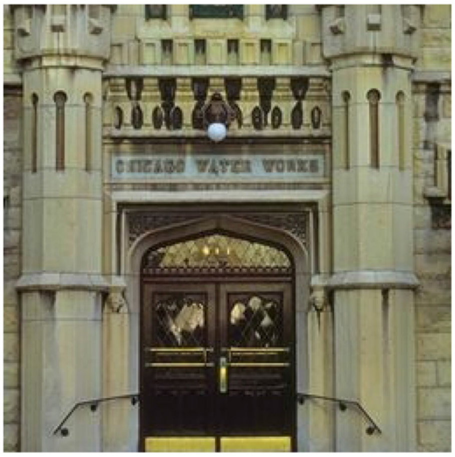


FIGURE 1 | An example of disagreement from LabelMe Ground truth label: *insidecity*, crowd annotations: [*insidecity*:1, *tallbuilding*:2].

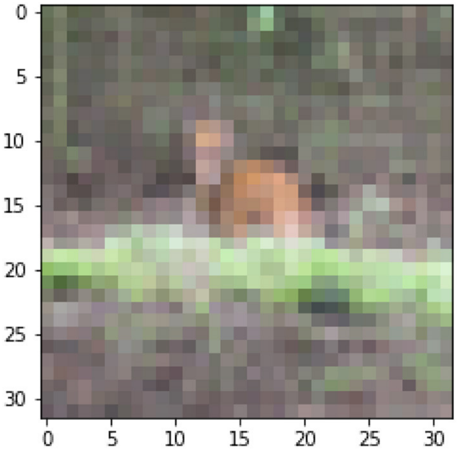


FIGURE 2 | An example of disagreement from CIFAR10H Ground truth label: *deer*, crowd annotations: [*dog*:33, *deer*:13, *horse*:4].

not possible to know which from the context. Here, *both* labels can be applied at the same time. Uma et al. (2021b) carried out an analysis of this dataset, finding that examples like **Figure 1** are prevalent. That is, the disagreement for this dataset is largely due to an **imprecise annotation scheme** where label categories are not necessarily mutually exclusive but may **overlap**. As a consequence, an annotator forced to choose one among the overlapping categories which apply to a particular image will likely make a random choice.

In our experiments, we randomly split the 10K images into training and test data (8,882 and 1,118 images respectively) to allow for ground truth and probabilistic evaluation. A total of 500 images from the dataset with gold labels were used as a development set.

3.2.4. The CIFAR-10H Corpus

As an example of a crowdsourced corpus containing very little disagreement and that primarily due to item difficulty, we used Krizhevsky's (2009) CIFAR-10H dataset, which consists of 60K tiny images from the web, carefully labeled, and expert-adjudicated to produce a single gold label for each image in one of 10 clearly distinct categories: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship,* and *truck*. Peterson et al. (2019) collected crowd annotations for 10K images from this dataset (the designated test portion) using Amazon Mechanical Turk, creating the CIFAR-10H dataset[12], which we use for our experiments.

The observed agreement for this dataset is 0.92, the highest among all the datasets. Clearly, the 2,457 annotators (about 51 annotators per item) found the annotation scheme to be clear and mostly agree with the expert opinion on what the label for each item would be. Notice that unlike in LABELME, there is no overlap: it is not possible for an object to belong to multiple categories. Cases of disagreement among annotators do occur, but they are primarily of the kind illustrated by **Figure 2**, which is because of the poor quality of the image, it is not possible to decide from the picture which animal is illustrated. Yet, there is no question that only one category can apply. We consider such cases as proper examples of **difficult to classify** items—items to which only one category from the scheme applies, yet problematic to classify because of noise.

We used the CIFAR-10H dataset for training and testing using a 70:30 random split, ensuring that the number of images per class remained balanced as in the original dataset. We also use a subset of Krizhevsky (2009) CIFAR-10 training dataset (3k images) as our development set.

3.2.5. The ArMIS Corpus

Finally, to exemplify an important source of disagreement—the fact that certain judgments are intrinsically subjective—we used our own ArMIS corpus (Almanea and Poesio, 2022). ArMIS is an Arabic misogyny dataset. It consists of 1K tweets each annotated with binary labels: 1 if the tweet expresses a misogynistic behavior according to the annotator's subjective point of view, 0 if the annotator believes that the tweet is not misogynistic. The tweets were collected using the Twitter API in October 2020, using a keywords list which was manually created specifically for this task, including specific slang words, phrases, and hashtags in order to get the related tweets, such as "Feminist," "Deficient mind and religion." The important aspect of this dataset is that it was annotated by three experts" annotators, carefully chosen to reflect different political views: liberal, moderate, and conservative. The annotators were asked to annotate the tweets based on their perspective.

The observed agreement of the annotators is 0.77, higher than the observed agreement of both POS and LABELME datasets (0.73), lower than the 0.92 observed agreement of CIFAR-10H, and equal to the observed agreement of PDIS. It is important to note while PDIS and ArMIS have the same level of disagreement, the nature and source of the disagreement for the ArMIS dataset

[12] https://github.com/jcpeterson/cifar-10h

differs from that of PDIS and indeed from the others. While Uma et al. (2021b) show that PDIS disagreements can be attributed to noise from spammers, the ambiguity of labels, or interface problems, an analysis of the disagreement in ArMIS showed the nature of the disagreements to be largely due to the **subjective viewpoints** of the diverse annotators.

For these experiments, we split the 964 tweets in ArMIS into 674 for training, 145 for validation, and 145 for testing. Gold labels were not obtained, as is fitting for a task of such as divisive nature, where annotator background plays a substantial role in how they label. However, as a compromise, we use majority voting to produce a hard label for hard evaluation purposes.

3.3. Base Models

The base models used in these experiments are the state-of-the-art or near state-of-art models used in previous work (Uma et al., 2020; Almanea and Poesio, 2022), many of which were made available to the participants to the 2021 SEMEVAL shared task on learning from disagreement (Uma et al., 2021a). We briefly summarize these models in this subsection.

3.3.1. The POS Tagging Model

For POS tagging, we used the bi-LSTM model (Plank et al., 2016) used by Uma et al. (2020). The model we used is improved from Plank et al. (2016) by using attention over the input token and character embeddings to learn contextualized token representations.

3.3.2. The PDIS Information Status Model

The model for this task was also developed by Uma et al. (2021a). Uma et al. combined the mention representation component of Lee et al.'s (2018) coreference resolution system with the mention sorting and non-syntactic feature extraction components of the IS classification model proposed by Hou (2016)[13] to create a novel IS classification model that outperforms (Hou, 2016) on the PDIS corpus. The training parameters were set following Lee et al. (2018).

3.3.3. The LabelMe Image Classification Model

For the LabelMe image classification, we replicated the model from Rodrigues and Pereira (2017). The images were encoded using pre-trained CNN layers of the VGG-16 deep neural network (Simonyan et al., 2013) and passed to a feed-forward neural network layer with a ReLU activated hidden layer with 128 units. A 0.2 dropout is applied to this learned representation which is then passed through a final layer with softmax activation to produce the model's predictions.

3.3.4. The CIFAR-10H-10 Image Classification Model

The trained model provided for this task is the ResNet-34A model (He et al., 2016), one of the best performing systems for the CIFAR-10 image classification. The publicly available Pytorch implementation of this ResNet model was used[14].

TABLE 1 | The effect of different loss functions for soft loss training on accuracy.

	POS	PDIS	LABELME	CIFAR-10H	ArMIS
SE Soft loss	79.20	92.90	84.21	63.49	76.83
CE Soft loss	79.80	92.86	84.66	66.54	**77.79**
KL Soft loss	**79.96**	92.86	84.73	**66.58**	76.41
Reverse KL Soft loss	79.81	**92.95**	**84.92**	63.71	75.59

The bold values indicate the best results for each model (indicated in the first column) using a given metric (indicated in the column header).

TABLE 2 | Results showing the accuracy (higher is better) and cross-entropy (lower is better) of soft loss models with and without temperature.

Task	Model	Accuracy↑	Cross-entropy ↓
LABELME	Reverse KL soft loss	84.97	1.671
LABELME	Reverse KL soft loss + T_i	**86.29***	**1.656**
POS	KL soft loss	79.96	**1.268***
POS	KL soft loss + T_i	**80.01**	1.547
PDIS	Reverse kl soft loss	92.95	0.467
PDIS	Reverse kl soft loss + T_i	**93.00**	**0.395***
CIFAR-10H	KL soft loss	**66.58***	**1.109***
CIFAR-10H	KL soft loss + T_i	63.89	1.223
ArMIS	CE soft loss	**77.79**	**0.586***
ArMIS	CE soft loss + T_i	76.83	0.636

An asterisk is used to indicate significantly better results.
The bold values indicate the best results for each model (indicated in the first column) using a given metric (indicated in the column header).

3.3.5. The ArMIS Arabic Misogyny Classification Model

For this task and dataset, we fine-tuned the state-of-the-art AraBERT base model (Antoun et al., 2020) with a maximum sequence length of 128, learning rate of 1e-5, batch size of 8, and training for 10 epochs.

4. RESULTS

Table 1 compares the effectiveness of different probability-comparing loss functions for making gold predictions, identifying the best soft loss function for each dataset. **Table 2** presents the results obtained for each task by models using the best soft loss function from **Table 1** with and without automatic temperature scaling, evaluated using both hard and soft metrics.

To account for non-deterministic model training effects, each model was trained and tested several times: (i) 30 times each for POS and LABELME (ii) 10 times each for PDIS, CIFAR-10H, and ArMIS owing to the complexity of the base models. We measure significance *via* bootstrap sampling, following Berg-Kirkpatrick et al. (2012) and Søgaard et al. (2014). The rest of this section discusses the results from these tables, highlighting significant results. The best result for each dataset is highlighted in bold.

4.1. Choosing the Loss Function

The aim of this preliminary experiment was to investigate Malinin and Gales's (2019) hypothesis that Reverse KL divergence is the most appropriate loss function for training models on

[13] This model was developed for fine-grained information status classification on the ISNOTES corpus (Markert et al., 2012; Hou et al., 2013).

[14] https://github.com/KellerJordan/ResNet-PyTorch-CIFAR10

datasets with high data uncertainty. We found that the Reverse KL soft loss function outperforms the other soft loss functions by a noticeable margin (0.19) for one dataset only, LABELME—though this margin is not significant[15]. This is the dataset for which we observe the most disagreement due to an annotation scheme with overlapping labels, as opposed to linguistic ambiguity (as in POS), or a combination of linguistic ambiguity and random noise (as in PDIS), or item difficulty (as in CIFAR-10H), or annotator biases (as in ArMIS). For CIFAR-10H, the dataset with the least amount of disagreement (and noise), as discussed by Uma et al. (2021b), we observe that Reverse KL soft loss falls nearly 3 significance points below either CE or KL soft loss. The SE loss function also performs poorly on this dataset, likely because SE optimizes the loss for non-modal classes, and this is an undesirable trait for a dataset like CIFAR-10H where the modal class is usually the gold class.

Following this experiment, we determine the best soft loss function for each dataset to be used as the starting point for the automatic temperature-scaled soft loss is as follows: CE for ArMIS, KL for POS and CIFAR-10H, and reverse KL for PDIS and LABELME.

4.2. Temperature Scaling Soft-Loss Learning

The first observation emerging from **Table 2** is that automatic temperature scaling only significantly improves results in one task: LABELME. In other words, our results would suggest that automatic temperature scaling only works when a disagreement arises from overlapping labels, resulting in the arbitrariness of ground truth.

In the next two datasets, POS and PDIS, the effect of temperature scaling on the performance of the models are mediocre or non-existent. These are the datasets for which we and Plank et al. (2014b) and Poesio et al. (2019) have shown that although a certain amount of noise is present, the disagreements are largely due to linguistic ambiguity and/or interface limitations.

At the other extreme, we have two datasets in which temperature scaling hurts performance. One of these is CIFAR-10H. This is a dataset with a very high observed agreement, 0.92. We also showed that the very few disagreements in this dataset are due to difficulty experienced by annotators when labeling blurry images. In other words, these disagreements are not systematic or a result of an imprecise annotation scheme but are due to the characteristics of the input. The other dataset for which automatic temperature scaling leads to a reduction in model performance is ArMIS. In this case, there is lower agreement than in CIFAR-10H, but this is not a reflection of systematic noise or data uncertainty, but of annotator uncertainty due to subjective biases.

5. INTERPRETING T_I

Our results show that among the datasets we considered in this study, automatic temperature scaling is effective for the one dataset in which disagreements are primarily due to what we may call **label arbitrariness**: the randomness in judgments originating from the fact that annotators have to choose one between multiple labels all of which could apply to an image and do so without appealing to any theory (given the vagueness of the annotation scheme). In this section, we examine the temperature predictions of the model for this dataset to understand what the model learns about label arbitrariness.

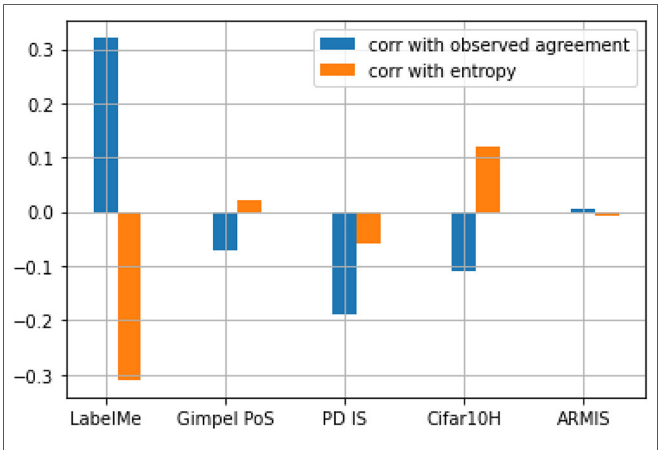


FIGURE 3 | Graph showing the correlation of T_i with observed agreement and entropy.

One way to do this is to measure the correlation of the temperature values to known measures of item agreement/uncertainty/difficulty. **Figure 3** shows the Pearson correlation (Pearson, 1896) between the temperature parameter and two such metrics of uncertainty/difficulty: observed agreement and normalized entropy. The results show that for LABELME, the only dataset for which our method produces a significant improvement over the soft-loss baseline, the model's T_i predictions have the strongest positive correlation to the observed agreement. This means that the model tended to make higher T_i predictions for items with a high observed agreement and lower T_i predictions for items with a low observed agreement. The model also has the strongest negative correlation to entropy. These two results suggest that for this dataset (but not for others), T_i is a moderately good predictor of uncertainty for this dataset as measured by observed agreement and entropy. What is it about the type of disagreement due to annotation schemes in which labels overlap that explains why temperature scaling improves performance with this kind of dataset, but not with others?

As mentioned earlier, in the one study of the differences between types of disagreement we are aware of, Reidsma and Carletta (2008) proposed a distinction between two types of disagreement between annotators and argued that they affect the performance of machine learning models in different ways. One kind is disagreements due to **random noise**, not conforming to any theorizable pattern. A second type is disagreements due to **bias**, which are identifiable through the occurrence of patterns of disagreement. The fact that automatic temperature scaling works best for disagreement due to overlap, which is the type of disagreement among those we studied that most

[15] Significance was computed using bootstrap sampling, following Berg-Kirkpatrick et al. (2012) and Søgaard et al. (2014).

resemble random noise because the annotators have to choose randomly; and it works worst for the clearest case of bias among our datasets, the misogyny data might suggest that automatic temperature scaling is a good method for adjusting model weights when the disagreements are due to random noise, but not when disagreement is due to bias. The mediocre results with PDIS and POS suggest that disagreements due to linguistic ambiguity sit somewhere in the middle, or do not fit this distinction at all. Of course, more research is needed to verify if this hypothesis also holds with other datasets in which disagreement is due to noise.

An alternative explanation can be found in the experiments conducted by Malinin and Gales (2019), who posit that overlapping labels (due to imprecise annotation schemes) introduce **data uncertainty**, resulting in multi-modal distributions[16]. The key characteristic of data uncertainty disagreement is that it is fully observable given the inputs and targets, without the need to appeal to linguistic theory (as in linguistic ambiguity) or annotator background (as in subjectivity disagreement). As such, a network of weights and biases (a machine annotator if you will), given the inputs and label distribution would also experience uncertainty predicting the targets for such images as human annotators do. In fact, an examination of the model's output distribution for the instances with the lowest temperature predictions shows that the model assigned the lowest temperatures (= highest uncertainty) to images belonging to the categories *tall building*, *street*, or *inside city*, the categories for which the annotators most disagree with the gold (**Figure 4** shows the class proportions of images 1st quartile range of temperature while **Figure 5** shows the confusion matrix between the majority and the gold). By calibrating its predictions by its level of certainty for each item, the model was able to fine-tune and improve its performance. Again, more research with other datasets characterized by data uncertainty will be required.

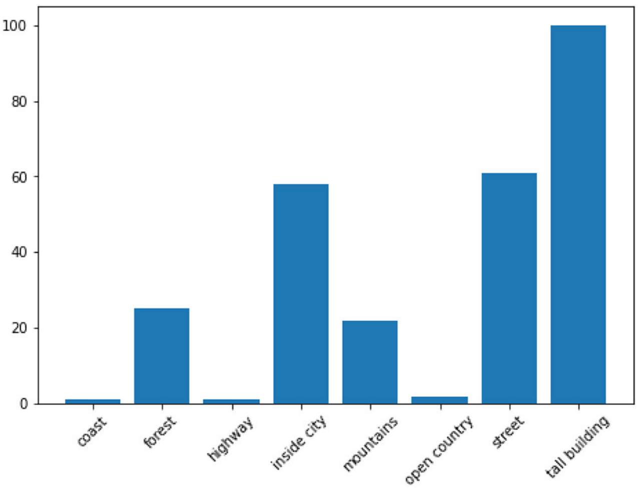


FIGURE 4 | Bar chart showing the gold label distribution of the images with the lowest temperature (images in the 1st quartile range of temperature), i.e., the lowest certainty.

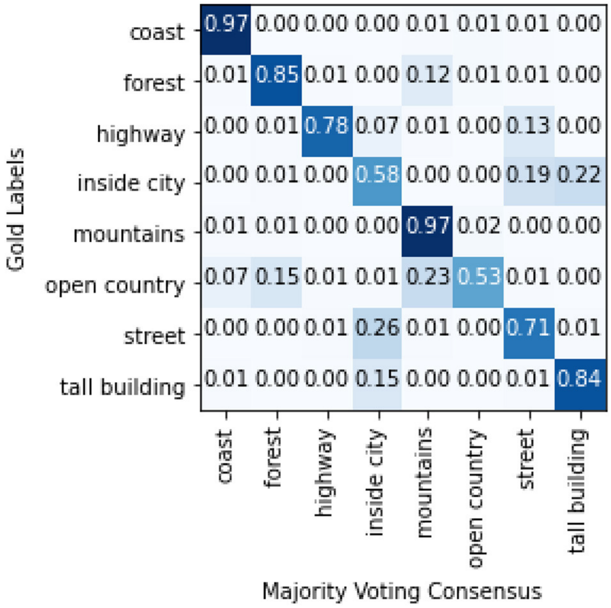


FIGURE 5 | Confusion matrix between gold labels and majority voting consensus for LabelMe.

6. CONCLUSION AND FUTURE WORK

Not all disagreements are the same, and it has been shown that not all approaches for learning from disagreement work equally well with datasets containing different types of disagreement (Uma et al., 2021b). In this paper, we reported on experiments on the use of automatic temperature scaling in a learning-from-disagreements setting as a way for automatically adjusting a model to take into account the peculiarities of a particular dataset. Our results show that model calibration *via* automatic temperature scaling can be a simple yet effective approach to improving model performance, particularly with learning ground truth predictions, but only with high disagreement datasets where the disagreements are due to overlapping labels.

We analyzed the temperature values of the successful model in a dataset of this type, to find that the temperature values have some correlation with two known measures of item disagreement/uncertainty—a positive correlation of about 0.3 with an observed agreement and a negative correlation of about 0.3 with entropy. We also observed that the model assigns the lowest temperature to instances with one of the three categories *inside city, street, tall building* shown by Uma et al. to be overlapping. We also found, however, that in datasets where disagreement is due to different reasons, the approach does not work so well.

We provide two possible explanations: automatic temperature scaling provides a good model of uncertainty when disagreements are due to random noise, but not when they are due to biases and automatic temperature scaling is a

[16] The results from **Table 1** while not significant do suggest that of the datasets examined in this work, LABELME has the most data uncertainty.

good indicator of data uncertainty. Further research is however needed to test these explanations with other datasets with the same characteristics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://zenodo.org/record/5130737#.YP_V9o5KiUk.

AUTHOR CONTRIBUTIONS

AU: conceptualization, methodology, software, formal analysis, investigation, visualization, and writing. DA: software, investigation, data curation, and writing. MP: conceptualization, methodology, writing—review and editing, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

AU and MP are supported by the DALI project, ERC Advanced Grant 695662 to MP. DA was supported by a studentship from the Saudi government.

ACKNOWLEDGMENTS

We thank Valerio Basile, Bob Carpenter, Tommaso Fornaciari, Dirk Hovy, Becky Passonneau, Silviu Paun, and Barbara Plank for many useful discussions and comments.

REFERENCES

Almanea, D., and Poesio, M. (2022). The ARMIS dataset of misogyny in arabic tweets. Submitted for publication.

Antoun, W., Baly, F., and Hajj, H. (2020). "AraBERT: transformer-based model for Arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, With a Shared Task on Offensive Language Detection* (Marseille: European Language Resource Association), 9–15.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.* 34, 555–596. doi: 10.1162/coli.07-034-R2

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., et al. (2021). "We need to consider disagreement in evaluation," in *BPPF* (Online).

Beigman, E., and Beigman Klebanov, B. (2009). "Learning with annotation noise," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Suntec: Association for Computational Linguistics), 280–287.

Beigman-Klebanov, B., and Beigman, E. (2009). From annotator agreement to noise models. *Comput. Linguist.* 35, 495–503. doi: 10.1162/coli.2009.35.4.35402

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). "An empirical investigation of statistical significance in NLP," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island: Association for Computational Linguistics), 995–1005.

Carpenter, B. (2008). *Multilevel Bayesian Models of Categorical Data Annotation*. Available online at: http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf

Dawid, A. P., and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *J. R. Stat. Soc. Ser. C* 28, 20–28.

Dumitrache, A., Aroyo, L., and Welty, C. (2018). "Crowdsourcing semantic label propagation in relation classification," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (Brussels: Association for Computational Linguistics), 16–21.

Dumitrache, A., Aroyo, L., and Welty, C. (2019). "A crowdsourced frame disambiguation corpus with ambiguity," in *Proceedings of North American Chapter of the Association for Computational Linguistics* (Minneapolis, MN).

Fleiss, J. L., Levin, B., and Paik, M. C. (2004). *The Measurement of Interrater Agreement*, 598–626. Sydney, NSW: John Wiley & Sons, Ltd. p. 598–626.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). "Beyond black & white: leveraging annotator disagreement via soft-label multi-task learning," in *Proceedings of North American Chapter of the Association for Computational Linguistics* (Online).

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., et al. (2011). "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, OR: Association for Computational Linguistics), 42–47.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW), 1321–1330.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (Amsterdam: IEEE), 770–778.

Hou, Y. (2016). "Incremental fine-grained information status classification using attention-based lstms," in *COLING* (Osaka).

Hou, Y., Markert, K., and Strube, M. (2013). "Global inference for bridging anaphora resolution," in *HLT-NAACL* (Seattle, WA).

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). "Learning whom to trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Atlanta, GA: Association for Computational Linguistics), 1120–1130.

Jamison, E., and Gurevych, I. (2015). "Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon: Association for Computational Linguistics), 291–297.

Krizhevsky, A. (2009). *Learning Multiple Layers of Features From Tiny Images*. Available online at: https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

Lee, K., He, L., and Zettlemoyer, L. (2018). "Higher-order coreference resolution with coarse-to-fine inference," in *NAACL-HLT* (New Orleans, LA).

Malinin, A., and Gales, M. (2019). "Reverse kl-divergence training of prior networks: improved uncertainty and adversarial robustness," in *NeurIPS* (Vancouver, BC).

Markert, K., Hou, Y., and Strube, M. (2012). "Collective classification for fine-grained information status," in *ACL* (Jeju).

Paun, S., Artstein, R., and Poesio, M. (2022). *Statistical Methods for Annotation Analysis*. Claypool, IN: Morgan.

Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., and Poesio, M. (2018). Comparing bayesian models of annotation. *Trans. Assoc. Comput. Linguist.* 6, 571–585. doi: 10.1162/tacl_a_00040

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lond. Ser. A* 187, 253–318.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. (2019). "Human uncertainty makes classification more robust," in *2019 IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 9616–9625.

Plank, B., Hovy, D., and Søgaard, A. (2014a). "Learning part-of-speech taggers with inter-annotator agreement loss," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg: Association for Computational Linguistics), 742–751.

Plank, B., Hovy, D., and Søgaard, A. (2014b). "Linguistically debatable or just plain wrong?," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Baltimore, MD: Association for Computational Linguistics), 507–511.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Berlin: Association for Computational Linguistics), 412–418.

Platt, J. C. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers* (Nebraska, NA: MIT Press), 61–74.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Intell. Interact. Syst.* 3, 1–44. doi: 10.1145/2448116.2448119

Poesio, M., Chamberlain, J., Paun, S., Yu, J., Uma, A., and Kruschwitz, U. (2019). "A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN: Association for Computational Linguistics), 1778–1789.

Poesio, M., Sturt, P., Arstein, R., and Filik, R. (2006). Underspecification and anaphora: theoretical issues and preliminary evidence. *Discourse Processes* 42, 157–175. doi: 10.1207/s15326950dp4202_4

Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., and Moy, L. (2010). "Learning from crowds," *J. Mach. Learn. Res.* 11, 1297–1322. Available online at: https://jmlr.org/papers/v11/raykar10a.bib

Recasens, M., Hovy, E., and Martí, M. A. (2011). Identity, non-identity, and near-identity: addressing the complexity of coreference. *Lingua* 121, 1138–1152. doi: 10.1016/j.lingua.2011.02.004

Reidsma, D., and Carletta, J. (2008). Reliability measurement without limits. *Comput. Linguist.* 34, 319–326. doi: 10.1162/coli.2008.34.3.319

Rodrigues, F., and Pereira, F. (2017). Deep learning from crowds. Available online at: https://www.bibsonomy.org/bibtex/1a9d85dd14f242883706b4b37e716429e/ghagerer

Russell, B., Torralba, A., Murphy, K., and Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *Int. J. Comput. Vision* 77, 157–173. doi: 10.1007/s11263-007-0090-8

Sheshadri, A., and Lease, M. (2013). "Square: a benchmark for research on computing crowd consensus," in *HCOMP* (Palm Springs).

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv[Preprint].arXiv:1312.6034.* doi: 10.48550/arXiv.1312.6034

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. (1994). "Inferring ground truth from subjective labelling of venus images," in *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94 (Cambridge, MA: MIT Press), 1085–1092.

Søgaard, A., Johannsen, A., Plank, B., Hovy, D., and Alonso, H. M. (2014). "What's in a p-value in nlp?," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (Baltimore, MD), 1–10.

Uma, A., Fornaciari, T., Dumitrache, A., Miller, T., Chamberlain, J. P., Plank, B., et al. (2021a). "Semeval-2021 task 12: learning with disagreements," in *SEMEVAL* (Online).

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2020). "A case for soft-loss functions," in *Proc. of HCOMP* (Online).

Uma, A., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021b). Learning with disagreements. *J. Art. Intell. Res.* 4, 201–213.

Versley, Y. (2008). Vagueness and referential ambiguity in a large-scale annotated corpus. *Res. Lang. Comput.* 6, 333–353. doi: 10.1007/s11168-008-9059-1

Whitehill, J., Fan Wu, T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (2009). "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Vancouver, BC: Curran Associates, Inc), 2035–2043.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 22 April 2022
doi: 10.3389/frai.2022.826499

EXP-Crowd: A Gamified Crowdsourcing Framework for Explainability

Andrea Tocchetti[1]*, *Lorenzo Corti*[1], *Marco Brambilla*[1] *and Irene Celino*[2]

[1] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy,* [2] *Cefriel, Milan, Italy*

The spread of AI and black-box machine learning models made it necessary to explain their behavior. Consequently, the research field of Explainable AI was born. The main objective of an Explainable AI system is to be understood by a human as the final beneficiary of the model. In our research, we frame the explainability problem from the crowds point of view and engage both users and AI researchers through a gamified crowdsourcing framework. We research whether it's possible to improve the crowds understanding of black-box models and the quality of the crowdsourced content by engaging users in a set of gamified activities through a gamified crowdsourcing framework named EXP-Crowd. While users engage in such activities, AI researchers organize and share AI- and explainability-related knowledge to educate users. We present the preliminary design of a game with a purpose (G.W.A.P.) to collect features describing real-world entities which can be used for explainability purposes. Future works will concretise and improve the current design of the framework to cover specific explainability-related needs.

Keywords: explainability, crowdsourcing, gamification, game with a purpose, Explainable AI

OPEN ACCESS

Edited by:
Ujwal Gadiraju,
Delft University of Technology,
Netherlands

Reviewed by:
Gaole He,
Delft University of Technology,
Netherlands
Burcu Sayin Gánel,
University of Trento, Italy

****Correspondence:***
Andrea Tocchetti
andrea.tocchetti@polimi.it

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence,
a section of the journal Frontiers in Artificial Intelligence

Received: *30 November 2021*
Accepted: *07 March 2022*
Published: *22 April 2022*

Citation:
Tocchetti A, Corti L, Brambilla M and Celino I (2022) EXP-Crowd: A Gamified Crowdsourcing Framework for Explainability.
Front. Artif. Intell. 5:826499.
doi: 10.3389/frai.2022.826499

1. INTRODUCTION

Over the last decades, the development of new Artificial Intelligence (AI) technologies brought forth the necessity of improving their understandability. In Explainable AI (XAI), most researchers develop algorithms to either explain models or improve their intrinsic explainability. The main problem associated with the understandability of an AI system is the gap between the explanation and the level of understanding of non-expert people. Such a gap is mainly influenced by the shape of the explanation (i.e., textual, visual, low-level details, etc.), its complexity, the persons level of knowledge, and many other factors associated with both the model and human side. In particular, while sometimes it is possible to re-shape the explanation to improve its understandability for non-experts, it is challenging to leverage people's knowledge as they are usually engaged in validation and data collection activities.

Alongside the development of AI systems, the need for training and labeled data has grown as well. Therefore, resorting to crowdsourcing has become essential to collect knowledge at scale. As such processes can sometimes be tedious and repetitive, researchers developed strategies to improve their design and effectiveness. In particular, Von Ahn (2006) proposed a *human computation paradigm* influencing the design of crowdsourcing activities, the so-called "Games With A Purpose" (G.W.A.P.). Such a paradigm enhances crowdsourcing endeavors through Gamification (Hamari, 2019), making them more entertaining for the people to partake.

Our research longs for envisioning an open gamified crowdsourcing framework with the final aim of (1) improving the capability of the crowd to understand black-box AI models explanations,

(2) improving the quality of the explanations provided by a black-box model by engaging the crowd to provide helpful content to AI practitioners, and (3) evaluating whether providing structured AI-related knowledge and engaging the crowd in explainability-related activities is an efficient way to achieve these objectives. As a first use case, our research covers image classification models. We explore user engagement, gamification, and knowledge collection and structuring to answer our research questions. Ultimately, we strive to create an open community through which users learn to understand the behavior of black-box models, therefore, providing value for both the developers and themselves.

The rest of this article is structured as follows. Section 2 provides an overview of explainability, crowdsourcing, and gamification. Section 3 outlines the preliminary framework design we envision, including a use case of a gamified activity for data collection and structuring and some use cases. Section 4 discusses the main advantages and limitations of the framework and explains how to overcome such restraints. A discussion on the gamified activity is also provided. Finally, Section 5 summarizes the critical contributions featured within this article and discusses the following research steps.

2. RELATED WORKS AND BACKGROUND

2.1. Explainability

One of the most well-known Artificial Intelligence (AI) branches is Machine Learning (ML). In ML, algorithms train models to perform predictions, classifications, groupings, and other tasks by learning from data. The development of Deep Learning (DL) and Deep Neural Networks (DNN) increased Machine Learning models' accuracy and performance at the expense of their interpretability. Indeed, most DNNs are referred to as "black-box" (or opaque) models. The input and output of a black-box model are known, while it is complex to understand its internal logic. They are opposed to "white-box" models, in which the internal logic is either known or easily understandable.

As of today, there is no unique definition of model explainability (Vilone and Longo, 2020). Despite the ongoing research efforts to define the fundamentals of an explainable AI system, most definitions are either domain- or problem-specific and are usually used interchangeably across different research fields (Guidotti et al., 2018). In the definitions provided by Barredo Arrieta et al. (2020), the notion of "human understandability" is the most important concept associated with Explainable AI. At the same time, other scholars consider different concepts depending on their research focus, like transparency (Belle and Papantonis, 2020) and explainability (Guidotti et al., 2018, Hu et al., 2021). In their definition of Explainable AI, (Barredo Arrieta et al., 2020) highlight that the understandability of an explanation is influenced by the ones to whom it is provided, i.e., *the audience*. In particular, depending on the person's knowledge about AI and ML, an explanation can be shaped differently. For example, an AI expert would probably prefer a detailed model description. On the other side, an inexperienced user would favor a small set of examples describing the system's behavior. Moreover, the authors state that an AI must generate an explanation *"clear or easy to understand,"* even though the concept of being easy to understand is not the same for everyone.

Regardless of the variety of explainability-related definitions provided in the literature, the researchers' community agreed that the main objective of an Explainable AI system is to be understood by a human as the final beneficiary of the model. Despite such an objective, XAI studies mainly approach the problem from a model-centric perspective rather than a user-centric one, overshadowing the level of users' understanding of the model. In particular, end-users and experts are frequently engaged in the later validation stages to evaluate the level of understandability of the model either directly or through simulated user experiments (Ribeiro et al., 2016; Lundberg and Lee, 2017).

2.2. Crowdsourcing and Gamification

Artificial Intelligence methods—especially Deep Learning approaches—require a large amount of high-quality data, whose collection is demanding and challenging. The widespread use of the internet allows researchers to engage virtually unlimited people to cover their data needs. Indeed, crowdsourcing has become common and encompasses academic studies and private companies' interests. Crowdsourcing can be defined as a participative online activity in which a group of individuals with varying features is engaged in undertaking a task as part of a process mutually benefiting participants and crowdsourcers (Estellés-Arolas and de Guevara, 2012). This methodology's advantages include lower costs, greater speed, and a higher degree of diversity by engaging a large and heterogeneous pool of people. This open-source practice either allows the collection of a wide variety of data, including peoples ideas and preferences (Balayn et al., 2021b), or the accomplishment of a task (e.g., labeling a large number of images) (Mishra and Rzeszotarski, 2021).

Sometimes, crowdsourcing is enhanced with gamification (Hamari, 2019) to make such a process more engaging, drive users' behaviors, and structure the collected data. Gamification uses people's motivations to achieve such objectives. Ryan and Deci (2000) accurately describe the influence of motivations on human decisions, mainly distinguishing between *intrinsic* and *extrinsic* motivations. Their definitions can be summarized as *"the motivation to perform a behavior or engage in an activity for our own sake rather than the desire for some external reward"* and *"the motivation to perform a behavior or engage in an activity due to a separable outcome"* (Lee et al., 2016), respectively. Following such a dichotomy, gamified approaches can be organized based on the kind of motivation they leverage. For example, pointification, leaderboards, etc. affects extrinsic motivation while receiving feedback (Hamari and Koivisto, 2013) and learning (Cerasoli et al., 2014) influence the intrinsic one. Moreover, an extrinsic-oriented design results in a good initial level of engagement, while it is necessary to apply an intrinsic-oriented design to achieve a long-lasting engagement (Rapp, 2015).

Gamification and G.W.A.P. have also been widely applied in computer science. Lu et al. (2021) developed a *Peek-a-boom*-based XAI evaluation, demonstrating the presence of differences between crowd-based and automatic assessment. Balayn et al. (2021a) developed *FindItOut*, a game with a purpose based on the *GuessWho* game with the final aim of collecting and organizing knowledge for researchers and AI practitioners. Speer et al. (2009) presented a gamified interface to acquire common sense knowledge through a *20 Questions* game which motivates contributions and improves the throughput of new knowledge. Other than contributing to data collection, it has also been demonstrated that Gamification can be effective in education and learning (Buckley and Doyle, 2016, Welbers et al., 2019). In particular, leveraging intrinsic motivation through feedback cycles is an effective way to enhance learning (Lee and Hammer, 2011).

3. METHODS

The main actors engaged within our explainability-oriented crowdsourcing framework fall into two categories: *users*, who get involved by playing gamified activities, and *AI practitioners/researchers*, who set up games and share knowledge about AI, ML, and explainability, since they exhibit a high level of understanding of these fields. **Figure 1** provides a simple overview of the interaction flow proposed within the framework.

The following sections describe each part of the framework and provide some use cases to clarify their structure. These will be mainly associated with the researcher side since most of the activities described for the user side are simple. We use a persona named "Bill" to represent our researcher. We will illustrate how he explores and interacts with the final implementation of our framework, i.e., a web-based platform.

3.1. Knowledge Assessment

As one of the main objectives of our framework is to improve the capability of the crowd to understand black-box models' explanations, educating users on AI-related topics is essential. Therefore, the first step is an assessment questionnaire through which their knowledge about AI and explainability will be assessed. Users will be asked to answer a series of multiple-choice questions. Depending on their results, they will be assigned a category representing their level of expertise. Users can improve their category by engaging with the proposed activities and enhancing their skills.

The research community will be requested to build a collection of multiple-choice questions employed in the assessment questionnaire and within the activities. Each question is made of (1) a set of texts through which the question is asked, (2) a set of correct answers, (3) the explanation associated with each correct answer, (4) a set of incorrect answers, (5) a difficulty score, and (6) a category. Questions must receive the approval of the community to guarantee the quality of the content provided. Therefore, each question must undergo a period of evaluation in which the community members can improve them by suggesting updates and proposing new answers and explanations. After this period, it is approved if the question received enough positive evaluations. Approved questions will be openly available to the whole research community as researchers may want to re-assess the users' knowledge as they engage with one of their activities. After a question is approved, researchers can still improve it by providing new content for elements (1–4).

***Use Case—Researcher**. Bill is a researcher who needs data about real-world entities for his research. When exploring the platform, Bill discovers a picture-based activity that would fit his needs. Even though he would like to set it up immediately, he also wants to evaluate the knowledge of the users who will perform his activity beforehand. Therefore, he explores the section dedicated to creating multiple-choice questions about AI, looking for questions that fit the context of his research. Unable to find questions that suit his needs, he submits new questions. A few days after his submission, he noticed that the researcher community proposed some improvements for the questions (e.g. by providing new answers). Bill approves a few of them. After a few more days, the question is approved.*

3.2. Education

Following the initial assessment, users will be schooled while engaging with the framework. In particular, knowledge will be provided in different shapes. The following list describes how knowledge about AI, ML, and explainability will be provided to users.

- **Questionnaire:** Researchers may set up a small quiz before their activity made of an arbitrary number of approved questions. For each question, they choose its text, the list of answers, and the explanation of the correct answer. Such a quiz would provide knowledge to users through the questions' explanations while allowing researchers to evaluate the level of education of the people playing the activity.
- **Knowledge Sharing:** Researchers can summarize, organize, and share knowledge by setting up tailored content for the users to read and study (i.e., the summary of a paper, the outline of the knowledge related to a specific AI topic, etc.). Each publication is made of a title, the topic it discusses, a brief description of the content, and the content itself. Researchers can also share scientific articles for the users to read. Only minimal information will be collected and shared like title, authors, and DOI. Users should access such articles by themselves.
- **Debating:** Researchers and users can discuss subjects of interest in a forum-like fashion. We argue that debating with knowledgeable people would improve the users' knowledge.

***Use Case—Researcher**. Bill would like users to understand how machine learning models learn so that users performing his activity can provide better inputs. Therefore, he collects knowledge from scientific documents, summarizes it, and shares it in the "Education" section. Bill achieves his first publication entitled "Understanding the way ML models learn from pictures: A simplified overview." He also provides a custom picture and a few references to the articles he used to write it within the publication. Bill reads an exciting article about his research topic a few days later. As it may improve the users' knowledge even further, he shares it by providing the necessary information.*

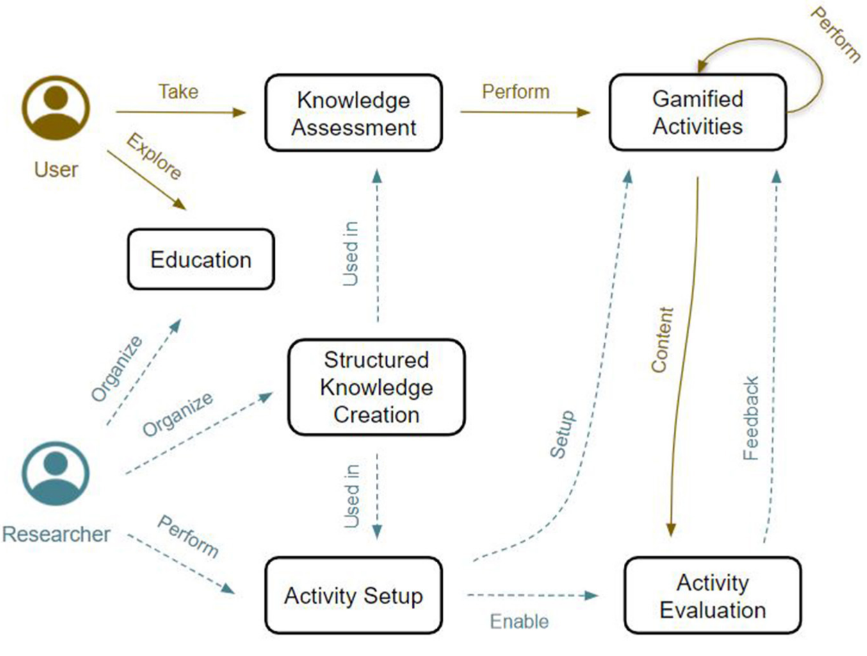


FIGURE 1 | Interaction flows of researchers (dashed cyan arrows) and users (orange plain arrows) with the activities devised within our framework, as described in Section 3. Researchers organize users' knowledge and set up activities to collect data. As users engage with such activities, they provide Content to researchers. In turn, researchers give the user feedback about the activity they performed. Such feedback aims to improve users' understanding of the activity itself, the knowledge and the context provided within it.

3.3. Gamification

Gamified activities are the core elements of our framework. The following sections discuss the steps a researcher must accomplish to set up and evaluate the outcomes of an activity.

Activity Setup. AI practitioners can pick between predefined activities and set up the necessary content depending on their needs. These activities range between data collection, explainability evaluation, etc.

Setting up an activity involves a set of passages, depending on the activity. In general, all setup processes share a questionnaire setup step, a context setup step, and an activity setup step. In the first setup step, researchers decide whether to include a Questionnaire (as described in "Education") and potentially organize its questions. In the second step, the researcher is asked to set up the content provided to the users to understand the context of their research, relevant concepts to know while carrying out the activity, etc. Finally, they have to provide all the necessary material to set up the actual activity. Practitioners can include additional control questions to the questionnaire and the actual activity to keep track of the user's level of attention. Practitioners can also select an advised knowledge level to provide an overview of the complexity of the concepts presented within the activity.

Use Case—Researcher. *Bill is finally ready to set up his first activity. In the questionnaire setup step, he picks the questions (including the ones he got approved before), their answers and their explanations. In the context setup step, he provides the context of its research, describing what it consists of. Bill also provides some of the content from the knowledge summary he shared for those who didn't read it. As the last step, he gives the pictures, labels, and necessary content for the picture-based activity.*

Activity Evaluation. Users are only asked to play gamified activities while researchers perform many different tasks regarding the gamified activities. In particular, they can visualize relevant statistics about the users that partook in the activity they set up, including the answers to the questionnaire (if present), the outcome of the activity, whether the user successfully answered the questions, the knowledge level of the users, etc. The role of the researcher in this final step is to evaluate the users and potentially provide feedback. They have to identify those users who stood out, like those who answered correctly to a high number of questions (compared to their level of knowledge), those who carried out a high-quality activity, etc. These users will be consequently awarded. In particular, these users will be awarded status-based awards that will make them distinguished community members.

Use Case—Researcher. *After a few weeks from publishing his activity, Bill overviews its outcomes. He notices that most users performed well while others outlined the pictures improperly. He picks the users who performed outstandingly and awards them. These users will be notified, and the award will be exhibited on their profiles. As one of the users answered most of the questions incorrectly and provided poor activity outcomes, Bill wrote them some advice on how to carry out the activity, also explaining some details related to how ML is applied in his research.*

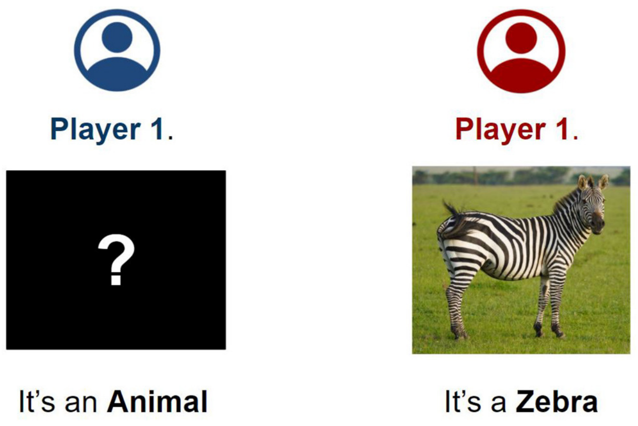


FIGURE 2 | The setup step of the gamified activity. *Player 1* is provided with the category of the entity they have to guess (in this case, they have to guess an animal). *Player 2* is supplied with a picture of the entity and its name (in this case, they are provided with the picture of a zebra).

3.4. Gamified Activity: A Case Study

Finally, we describe a case study on image classification and understanding, which we use as proof of concept of a gamified activity to collect data to be employed in the field of explainable AI.

When addressing the explainability of image classification models, the crowd is usually engaged to highlight, label, and detail pictures. We assert that the outcome of such a task strictly depends on the images supplied, i.e., a person describing different pictures of the same entity may provide different details. In particular, we argue that the description of a subject, provided its picture, may be limited to or by the features displayed. Therefore, we claim it would be possible to improve the collected features by unbinding the images from the process since the person won't be limited by the representation of the entity they describe. In particular, we would like to answer to the following questions

- **(Q1) Is the picture displayed to the annotator causing bias when asked to describe the entity in the image?**
- **(Q2) Are we able to collect more features with respect to the standard annotation methods?**

Therefore, we design and evaluate the effectiveness of a Game With A Purpose (G.W.A.P.) to collect knowledge in terms of relevant features and descriptions of the analyzed content. Such features are organized in three categories, namely "abstract" (identified with "A,") "not represented in the picture" (identified with "NR,") and "represented in the picture" (identified with "R.") "R" features and "NR" features both represents "concrete features."

Inspired by Ahn et al. (2006), we designed a gamified activity where a pair of people play a guessing game. The game involves the following steps.

- **Initial Setup** step (**Figure 2**). *Player 1* is provided with the entity category they have to guess. *Player 2* is shown the picture of the entity and its name.
- **Basic Turn** (**Figure 3**, on the left). *Player 1* asks closed questions about the features of the entity to guess. *Player 2* answers the questions. *Player 1* may either ask questions freely or fill in predefined question templates (i.e., "Does it have ...?," "Does it ...?," etc.). If the answer is affirmative, *Player 2* is asked to carry out the **Annotation Step**.
- **Annotation Step** (**Figure 3**, on the right). Whenever the answer to a question is affirmative, *Player 2* is asked to perform a series of simple tasks to identify the guessed feature in the picture they were provided with, if possible. First, they are asked whether the element is displayed in the image. If so, they are requested to outline them in the picture. Otherwise, they are asked whether the feature is an abstract one.
- **Hint Step**. If *Player 1* guessed no features of the unknown entity in the last few questions, *Player 2* provides a bit of advice by providing a feature of the entity to *Player 1*. If possible, *Player 2* should provide a feature that *Player 1* already tried to guess. Therefore, *Player 1* will be able to proceed with the activity. *Player 2* is still required to carry out the **Annotation Step** for the hinted feature as it will still be considered in the final set of features.
- **Game Conclusion**. Finally, after *Player 1* has collected enough clues on the entity they are trying to guess, they can provide their final answer. If the answer is correct, the game is over; otherwise, the game moves on. When the game ends, *Player 1* is shown both the original picture and the ones with the outlined features to check that *Player 2* performed their task correctly. If any element has been improperly outlined or any question has been incorrectly answered, *Player 1* can provide their solution (i.e., answer and annotation). Such an action generates a conflict the researcher will resolve when the outcomes of the activity are provided.

Such an activity can be set up to have players mainly focus their questions on concrete features, abstract features or both. Moreover, such a gamified activity could be extended by applying the following changes, enhancing various steps of the activity:

- It would be possible to introduce a further step at the end of the activity where *Player 2* provides an additional picture of the same entity and outline the missing features on the new image.
- It would also be possible to introduce a further **Annotation Step** for *Player 1* at the end of the game to improve the reliability of the results, allowing the comparison of both players' annotation to identify inconsistencies in the provided outcomes.

4. PRELIMINARY EVALUATION

In this section, we report on a preliminary study on the effectiveness and impact of our approach. The experiments have been performed by selecting one entity category and by asking participants to interact over it. In particular, we picked "animals" as a category. We selected parrots and crocodiles as relevant representatives, and we collected a picture from Google Images

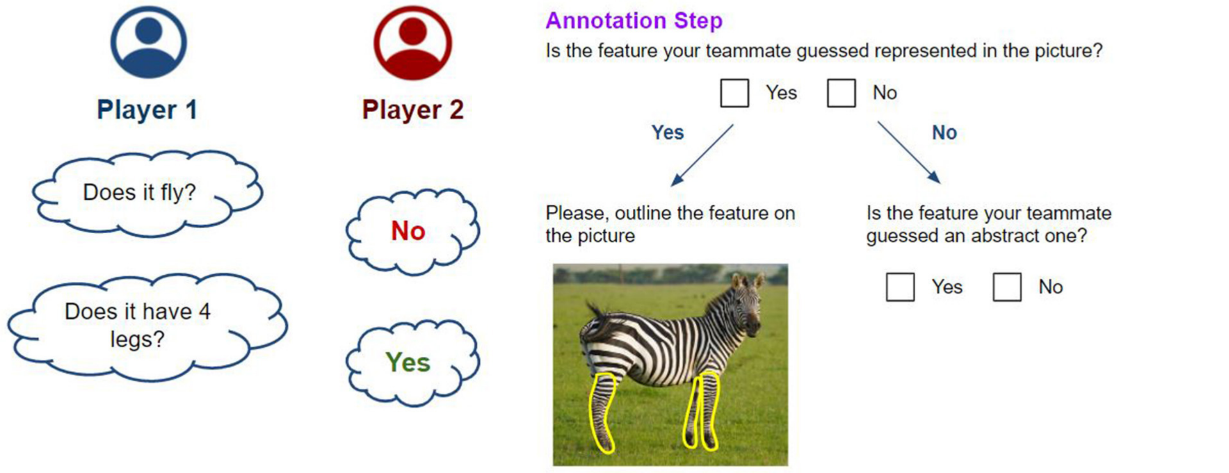

Player 1
Player 2
Does it fly?
No
Does it have 4 legs?
Yes
Annotation Step
Is the feature your teammate guessed represented in the picture?
□ Yes □ No
Yes
No
Please, outline the feature on the picture
Is the feature your teammate guessed an abstract one?
□ Yes □ No

FIGURE 3 | On the left, the Basic Turn of the gamified activity is displayed. *Player 1* asks yes or no questions about the entity. *Player 2* answers such questions. On the right, the Annotation Step is summarized. *Player 2* is asked to complete a series of simple tasks to identify the guessed feature by answering questions and potentially annotating the picture.

for each of them. We purposely selected an image partially representing the crocodile (i.e., only its head was visible) and a complete one for the parrot. We engaged 30 people aged between 24 and 30, mostly (60%) employed in IT-related sectors. Most of them (75%) achieved an educational level superior or equal to a bachelor's degree. The participants were randomly organized into three groups:

- The "annotation" group (comprising 6 people), focusing on outlining features on images;
- The "gamified activity (concrete)" group (comprising 12 people) focusing their questions on concrete features (i.e., "R" and "NR" features);
- And the "gamified activity (generic)" group (comprising 12 people), where members were allowed to ask questions about any features.

Depending on such a division, each person was provided with a document describing their activity. The members of each of the gamified activity groups have been internally organized in pairs to carry out the game, thus generating 6 pairs per group. Each player was given one picture to play with. Players were asked to follow the same procedure described in subsection 3.4, depending on their role and group. Each member of the pairs alternately played both roles. Overall, each group carried out 12 matches, (i.e., 6 matches per picture). Additionally, we asked people to keep track of each question and answer when playing as *Player 1*, and keep track of the suggestions provided when playing as *Player 2*. On the other hand, each of the 6 members of the "annotation" group was given two documents containing the chosen figures. They were appointed to describe the represented animal by providing a clear and short description of their features, its possible outline on the image, and its category.

5. RESULTS AND DISCUSSION

5.1. Gamified Activity

Following the preliminary experiment, we discuss the outcomes and the feedback we collected, concerning the research questions we wanted to address.

With respect to **(Q1)**, aiming at assessing the role of the specific picture used in generating bias in the player describing the displayed object, we observed that (as expected) most of the concrete details reported by each "annotation" group member were represented in the picture, 73% for the crocodile and 97% for the parrot (**Table 1A**). Within the same group, we outlined a clear tendency to report "R" features first and forget about features not represented within the picture. Indeed, 50% of the participants provided no "NR" features for the partial image. These observations are aligned with our initial thoughts and expectations. When a person is asked to describe an entity, it mainly attains to the particular representation provided in the picture rather than the actual entity, even when it is well-known. Moreover, we observed a significant difference in the ratio between the amount of "NR" and concrete features collected for the partial picture among the different experiments. In particular, such proportion grew from 27% in the "Annotation" task to 34% in the "Gamified Activity (concrete)." Such a difference is even more emphasized in the "Gamified Activity (general)" experiment. We also identify a 50% increase in the total amount of "NR" features collected by the "Gamified Activity (concrete)" group with respect to the "annotation" one. Therefore, we may argue that creating a sharp separation of roles and hiding the picture from the gamified activity contributes to reducing the bias it induces.

Regarding **(Q2)**, we argue that our methodology is able to identify more features w.r.t. a state of the art annotation method.

TABLE 1 | The table represents the average and the sample m.s.e. per participant for each feature type and for each picture.

(A) “Annotation” Group			
Picture	“R” Features	“NR” Features	“A” Features
Crocodile	3.6.7 ± 0.51	1.33 ± 1.63	1.5 ± 1.38
Parrot	5 ± 2	0.17 ± 0.48	2.17 ± 1.72
(B) “Gamified Activity (concrete)” Group			
Picture	“R” Features	“NR” Features	“A” Features
Crocodile	3.83 ± 1.94	2 ± 0.63	0.17 ± 0.41
Parrot	6 ± 0.89	0 ± 0	0.83 ± 0.41
(C) “Gamified Activity (generic)” Group			
Picture	“R” Features	“NR” Features	“A” Features
Crocodile	0.5 ± 0.55	1.17 ± 0.75	3.33 ± 0.81
Parrot	1.5 ± 0.55	0 ± 0	2.67 ± 1.51

The table is organized depending on the groups described in Section 4.

Indeed, when the participants were asked to focus on concrete features (**Table 1B**), we observed a 20% increase in the number of “R” features for the picture of the parrot and a 33% increase in the number of “NR” features for the crocodile one, with respect to the features identified by the “annotation” group by using traditional methods. When analysing the outcomes of the “gamified activity (general)” group, we identified a clear tendency to ask questions about abstract features (e.g., “Is it carnivorous?,” “is it oviparous?,” “Does it live in the Jungle?,” “Is it able to speak?,” etc.) resulting in a 55% increase of abstract features collected with respect to the “Annotation” task (**Table 1C**). We believe such a behavior is strictly related to humans’ capability to abstract concrete concepts and distinguish similar entities through peculiar and selective features, which (sometimes) are abstract. Questions on such selective features even played a fundamental role in the “Gamified Activity (concrete)” group, in which most people who had already collected a lot of concrete features, at the end of the process expressed the need to ask a few abstract questions to consolidate and finalize the identification of the animal. Furthermore, we believe that several descriptive dimensions, e.g., the selectivity of the features, and the category of the entity affect such behaviors.

We also collected some comments from the participants, whose feedback would lead to a significant improvement of the gamified activity. In particular, the following changes could be applied

- *Player 2* won’t provide annotations for the collected features during the activity but only at the end. Such a change would smooth the flow of the activity, making it quicker and even more enjoyable for both players.
- At the end of the activity, both *Player 1* and *Player 2* will carry out the **Annotation Step**, improving the consistency of the results and the amount of data collected.
- At the end of the activity, both players will be shown the picture of the entity to further enrich the collection of the features they already identified by describing those they can derive from the entity’s image.

In conclusion, we argue that our methodology extends gamified visual annotation and labeling methods, like the ones proposed in Runge et al. (2015) and Balayn et al. (2021a), mitigating the bias caused by pictures by hiding them, allowing an even more complete collection of features. Furthermore, our methodology can be easily extended by introducing further rules to shape and enhance its outcomes. Such an activity can be employed to collect data about what the model should know or should have learned about the entity. Such knowledge can be compared with the outcomes of other explainability methods to evaluate the difference between what the model knows and what it should know. Such a comparison can be carried out both for models learning from pictures of the entity - by comparing the heat maps derived from the model and the annotated “R” (and optionally “NR”) features—and textual descriptions of the entity—comparing the outcomes of saliency-based analyses and the collected features. Moreover, the collected knowledge could be further combined to enhance the outcomes of non-textual, local explainability methods or improve the textual description of textual ones. In particular, non-abstract features annotated by the crowd would be useful to describe pictures in which the same feature is detected by other methods (e.g., heat maps, etc.), while abstract details would be useful to complete textual descriptions, making them more human-understandable and human-like.

5.2. Framework

We argue that our framework would facilitate and structure the exchange of knowledge between the research community and the crowd, leading to an overall improvement of the content provided and the level of understanding of the engaged users. Moreover, the presented crowdsourcing framework engages the users on a different level with respect to other platforms mainly based on extrinsic rewards. In particular, user education would improve users’ awareness of what kind of knowledge an AI system needs, learns, and produce, enhancing their efficiency and shaping their mindset. Such a statement would also be amplified when a long-term engagement of the users is achieved.

We are aware of the limitations implied by our framework, namely the initial engagement gap, the necessity of keeping the users and the researchers engaged, and the high level of flexibility required to cover all the explainability-related aspects. Gamification will be helpful to compensate for the first two aspects, while the last one will be covered through an accurate design of the proposed activities. In particular, the design will include both extrinsic and intrinsic design elements to account for both the initial and long-term engagement, respectively. In particular, users’ side extrinsic design elements will consist of points, activity leaderboards, achievements (i.e., status as a reward), etc. Intrinsic design elements will be mainly associated with the education aspect as it is strictly related to one of the three innate psychological needs (Ryan and Deci, 2000), namely Competence (i.e., people are wishful to learn new skills and mastery tasks). On the other hand, we expect researchers to be engaged as they trade their scientific knowledge for data for their research. Moreover, developing a cooperative framework is challenging, especially when users and researchers must be engaged. We plan to engage users using renowned crowdsourcing platforms for testing purposes, while the initial engagement on the final release will be performed through the university and researchers' network.

6. CONCLUSIONS AND FUTURE WORKS

We presented the preliminary design of a crowdsourcing framework to create a cooperative cycle in which the crowd is taught about explainability-related topics and provides valuable content to AI practitioners. Gamification is applied to empower engagement and drive user behavior. The design and the preliminary evaluation of a gamified data collection activity is also provided. We argue that our research would improve the quality of the data collected to evaluate and enhance the explainability of black-box models. Future work will involve the improving of the design of both the presented activity—following the discussed changes—and the framework. We plan to execute further experiments to generalize the results on effectiveness and efficiency of our method, and to release an opensource crowdsourcing platform, which may be adopted by the broader research community.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

This research is partially supported by the European Commission under the H2020 framework, within project 822735 TRIGGER (TRends in Global Governance and Europe's Role) and by the contribution of the Ph.D. Scholarship on Explainable AI funded by Cefriel.

REFERENCES

Ahn, L. V., Kedia, M., and Blum, M. (2006). "'Verbosity: A game for collecting common-sense facts," in *In Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems, volume 1 of Games* (New York, NY: ACM Press), 75–78.

Balayn, A., He, G., Hu, A., Yang, J., and Gadiraju, U. (2021a). Finditout: A multiplayer gwap for collecting plural knowledge. In *Vol. 9 (2021): Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing* (Virtual), 190.

Balayn, A., Soilis, P., Lofi, C., Yang, J., and Bozzon, A. (2021b). "What do you mean? interpreting image classification with crowdsourced concept extraction and analysis," in *Proceedings of the Web Conference 2021 WWW '21* (New York, NY: Association for Computing Machinery), 1937–1948.

Barredo Arrieta, A., D'ıaz-Rodr'ıguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Belle, V., and Papantonis, I. (2020). Principles and practice of explainable machine learning. *CoRR*, abs/2009.11698.

Buckley, P., and Doyle, E. (2016). Gamification and student motivation. *Interact. Learn. Environ.* 24, 1162–1175. doi: 10.1080/10494820.2014.964263

Cerasoli, C., Nicklin, J., and Ford, M. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychol. Bull.* 140, 980–1008. doi: 10.1037/a0035661

Estellés-Arolas, E., and de Guevara, F. G.-L. (2012). Towards an integrated crowdsourcing definition. *J. Inf. Sci.* 38, 189–200. doi: 10.1177/0165551512437638

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 1–42. doi: 10.1145/3236009

Hamari, J. (2019). Gamification. *The Blackwell Encyclopedia of Sociology*. John Wiley and Sons, Ltd., 1–3. doi: 10.1002/9781405165518.wbeos1321

Hamari, J. and Koivisto, J. (2013). Social motivations to use gamification: An empirical study of gamifying exercise. *ECIS 2013 - Proceedings of the 21st European Conference on Information Systems*.

Hu, Z. F., Kuflik, T., Mocanu, I. G., Najafian, S., and Shulner Tal, A. (2021). "Recent studies of xai - review," in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization UMAP '21* (New York, NY: Association for Computing Machinery).

Lee, J., and Hammer, J. (2011). Gamification in education: what, how, why bother? *Acad. Exchange Quarter.* 15, 1–5.

Lee, W., Reeve, J., Xue, Y., and Xiong, J. (2016). Neural differences between intrinsic reasons for doing versus extrinsic reasons for doing: An fMRI study. *Neurosci. Res.* 68–72. doi: 10.1016/j.neures.2012.02.010

Lu, X., Tolmachev, A., Yamamoto, T., Takeuchi, K., Okajima, S., Takebayashi, T., et al. (2021). Crowdsourcing evaluation of saliency-based XAI methods. *CoRR*, abs/2107.00456.

Lundberg, S. M., and Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Mishra, S., and Rzeszotarski, J. M. (2021). Crowdsourcing and evaluating concept-driven explanations of machine learning models. *Proc. ACM Hum.-Comput. Interact.* 5, 1–26. doi: 10.1145/3449213

Rapp, A. (2015). A qualitative investigation of gamification: motivational factors in online gamified services and applications. *Int. J. Technol. Hum. Interact.* 11, 67–82. doi: 10.4018/ijthi.2015010105

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "'why should i trust you?': explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16* (New York, NY: Association for Computing Machinery), 1135–1144.

Runge, N., Wenig, D., Zitzmann, D., and Malaka, R. (2015). "Tags you don't forget: gamified tagging of personal images," in *14th International Conference on Entertainment Computing (ICEC)* (Trondheim), 301–314.

Ryan, R. M., and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037//0003-066x.55.1.68

Speer, R., Krishnamurthy, J., Havasi, C., Smith, D. A., Lieberman, H., and Arnold, K. C. (2009). "An interface for targeted collection of common sense knowledge using a mixture model," in *Proceedings of the 14th International Conference on Intelligent User Interfaces* (New York, NY).

Vilone, G., and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *CoRR*, abs/2006.00093.

Von Ahn, L. (2006). Games with a purpose. *Computer* 39, 92–94. doi: 10.1109/MC.2006.196

Welbers, K., Konijn, E. A., Burgers, C., de Vaate, A. B., Eden, A., and Brugman, B. C. (2019). Gamification as a tool for engaging student learning: A field experiment with a gamified app. *E-Learn. Digit. Media* 16, 92–109. doi: 10.1177/2042753018818342

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 18 May 2022
doi: 10.3389/frai.2022.828187

In Search of Ambiguity: A Three-Stage Workflow Design to Clarify Annotation Guidelines for Crowd Workers

Vivek Krishna Pradhan[1], *Mike Schaekermann*[2] *and Matthew Lease*[1,2]*

[1] University of Texas at Austin, Austin, TX, United States, [2] Amazon, Seattle, WA, United States

We propose a novel three-stage FIND-RESOLVE-LABEL workflow for crowdsourced annotation to reduce ambiguity in task instructions and, thus, improve annotation quality. Stage 1 (FIND) asks the crowd to find examples whose correct label seems ambiguous given task instructions. Workers are also asked to provide a short tag that describes the ambiguous concept embodied by the specific instance found. We compare collaborative vs. non-collaborative designs for this stage. In Stage 2 (RESOLVE), the requester selects one or more of these ambiguous examples to label (resolving ambiguity). The new label(s) are automatically injected back into task instructions in order to improve clarity. Finally, in Stage 3 (LABEL), workers perform the actual annotation using the revised guidelines with clarifying examples. We compare three designs using these examples: examples only, tags only, or both. We report image labeling experiments over six task designs using Amazon's Mechanical Turk. Results show improved annotation accuracy and further insights regarding effective design for crowdsourced annotation tasks.

Keywords: crowdsourcing, annotation, labeling, guidelines, ambiguity, clarification, machine learning, artificial intelligence

Edited by:
Fabio Casati,
Servicenow, United States

Reviewed by:
Hend Al-Khalifa,
King Saud University, Saudi Arabia
Marcos Baez,
Independent Researcher, Osnabrück, Germany

****Correspondence:***
Matthew Lease
ml@utexas.edu

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence,
a section of the journal Frontiers in Artificial Intelligence

Received: *03 December 2021*
Accepted: *18 March 2022*
Published: *18 May 2022*

Citation:
Pradhan VK, Schaekermann M and Lease M (2022) In Search of Ambiguity: A Three-Stage Workflow Design to Clarify Annotation Guidelines for Crowd Workers. Front. Artif. Intell. 5:828187. doi: 10.3389/frai.2022.828187

1. INTRODUCTION

While crowdsourcing now enables labeled data to be obtained more quickly, cheaply, and easily than ever before (Snow et al., 2008; Sorokin and Forsyth, 2008; Alonso, 2015), ensuring data quality remains something of an art, challenge, and perpetual risk. Consider a typical workflow for annotating data on Amazon Mechanical Turk (MTurk): a *requester* designs an annotation task, asks multiple workers to complete it, and then post-processes labels to induce final consensus labels. Because the annotation work itself is largely opaque, with only submitted labels being observable, the requester typically has little insight into what if any problems workers encounter during annotation. While statistical aggregation (Hung et al., 2013; Sheshadri and Lease, 2013; Zheng et al., 2017) and multi-pass iterative refinement (Little et al., 2010a; Goto et al., 2016) methods can be employed to further improve initial labels, there are limits to what can be achieved by *post-hoc* refinement following label collection. If initial labels are poor because many workers were confused by incomplete, unclear, or ambiguous task instructions, there is a significant risk of "garbage in equals garbage out" (Vidgen and Derczynski, 2020).

In contrast, consider a more traditional annotation workflow involving trusted annotators, such as practiced by the Linguistic Data Consortium (LDC) (Griffitt and Strassel, 2016). Once preliminary annotation guidelines are developed, an iterative process ensues in which: (1) a subset of data is labeled based on current guidelines; (2) annotators review corner cases and disagreements, review relevant guidelines, and reach a consensus on appropriate resolutions; (3) annotation guidelines are updated; and (4) the process repeats. In comparison to the simple crowdsourcing workflow above, this traditional workflow iteratively debugs and refines task guidelines for clarity and completeness in order to deliver higher quality annotations. However, it comes at the cost of more overhead, with a heavier process involving open-ended interactions with trusted annotators. Could we somehow combine these for the best of both worlds?

In this study, we propose a novel three-stage FIND-RESOLVE-LABEL design pattern for crowdsourced annotation which strikes a middle-ground between the efficient crowdsourcing workflow on one hand and the high quality LDC-style workflow on the other. Similar to prior study (Gaikwad et al., 2017; Bragg and Weld, 2018; Manam and Quinn, 2018), we seek to design a light-weight process for engaging the workers themselves to help debug and clarify the annotation guidelines. However, existing approaches typically intervene in a reactive manner *after* the annotation process has started, or tend to be constrained to a specific dataset or refinement of textual instruction only. By contrast, our approach is proactive and open-ended. It leverages crowd workers' unconstrained creativity and intelligence to identify ambiguous examples through an Internet search on the Internet and enriches task instructions with these concrete examples proactively upfront before the annotation process commences. Overall, we envision a partnership between the requester and workers in which each party has complementary strengths and responsibilities in the annotation process, and we seek to maximize the relative strengths of each party to ensure data quality while preserving efficiency.

Figure 1 depicts our overall workflow. In Stage 1 (FIND), workers are shown initial guidelines for an annotation task and asked to search for data instances that appear ambiguous given the guidelines. For each instance workers find, they are also asked to provide a short tag that describes the concept embodied by the specific instance which is ambiguous given the guidelines. Next, in Stage 2 (RESOLVE), the requester selects one or more of the ambiguous instances to label as exemplars. Those instances and their tags are then automatically injected back into the annotation guidelines in order to improve clarity. Finally, in Stage 3 (LABEL), workers perform the actual annotation using the revised guidelines with clarifying examples. The requester can run the LABEL stage on a sample of data, assess label quality, and then decide how to proceed. If quality is sufficient, the remaining data can simply be labeled according to the guidelines. Otherwise, Stages 1 and 2 can be iterated in order to further refine the guidelines.

To evaluate our three-stage task design, we construct six different image labeling tasks with different levels of difficulty and intuitiveness. We construct a test dataset that contains different ambiguous and unambiguous concepts. Starting from simple and possibly ambiguous task instructions, we then improve instructions *via* our three-stage workflow. Given expert (gold) labels for our dataset for each of the six tasks, we can evaluate how well-revised instructions compare to original instructions by measuring the accuracy of the labels obtained from the crowd.

1.1. Contributions

We provide initial evidence suggesting that the crowd can find and provide useful ambiguous examples which can be used to further clarify task instructions and that these examples may have the potential to be utilized to improve annotation accuracy. Our experiments further seem to suggest that workers can perform better when shown key ambiguous examples as opposed to randomly chosen examples. Finally, we provide an analysis of workers' performance for different intents of the same classification task and different concepts of ambiguity within each intent.

Our article is organized as follows. Section 2 presents Motivation and Background. Next, section 3 details our 3-Stage FIND-RESOLVE-LABEL workflow. Next, section 4 explains our experimental setup. Section 5 then presents the results. Finally, section 6 discusses conclusions and future directions.

2. MOTIVATION AND BACKGROUND

Consider the task of labeling images for object detection. For example, on MTurk one might post a task such as, "Is there a dog in this image?" Such a task appears to be quite simple, but is it? For example, is a wolf a dog? What about more exotic and unusual wild breeds of dogs? Does the dog need to be a real animal or merely a depiction of one? What about a museum model of an ancient but extinct dog breed, or a realistic wax sculpture What if the dog is only partially visible in the image? Ultimately, what is it that the requester really wants? For example, a requester interested in anything and everything dog-related might have very liberal inclusion criteria. On the other hand, a requester training a self-driving car might only care about animals to be avoided, while someone training a product search engine for an e-commerce site might want to include dog-style children's toys (Kulesza et al., 2014).

As this seemingly simple example illustrates, annotation tasks that seem straightforward to a requester may in practice embody a variety of subtle nuances and ambiguities to be resolved. Such ambiguities can arise for many reasons. The requester may have been overly terse or rushed in posting a task. They may believe the task is obvious and that no further explanation should be needed. They likely also have their own implicit biases (of which they may be unaware) that provide a different internal conception of the task than others might have. For example, the requester might be ignorant of the domain (e.g., is a wolf a type of dog?) or have not fully defined what they are looking for. For example, in information retrieval, users' own conception and understanding of what they are looking for often evolve during the process of search and browsing (Cole, 2011). We describe our own experiences with this in section 5.1.1. Annotators, on the other

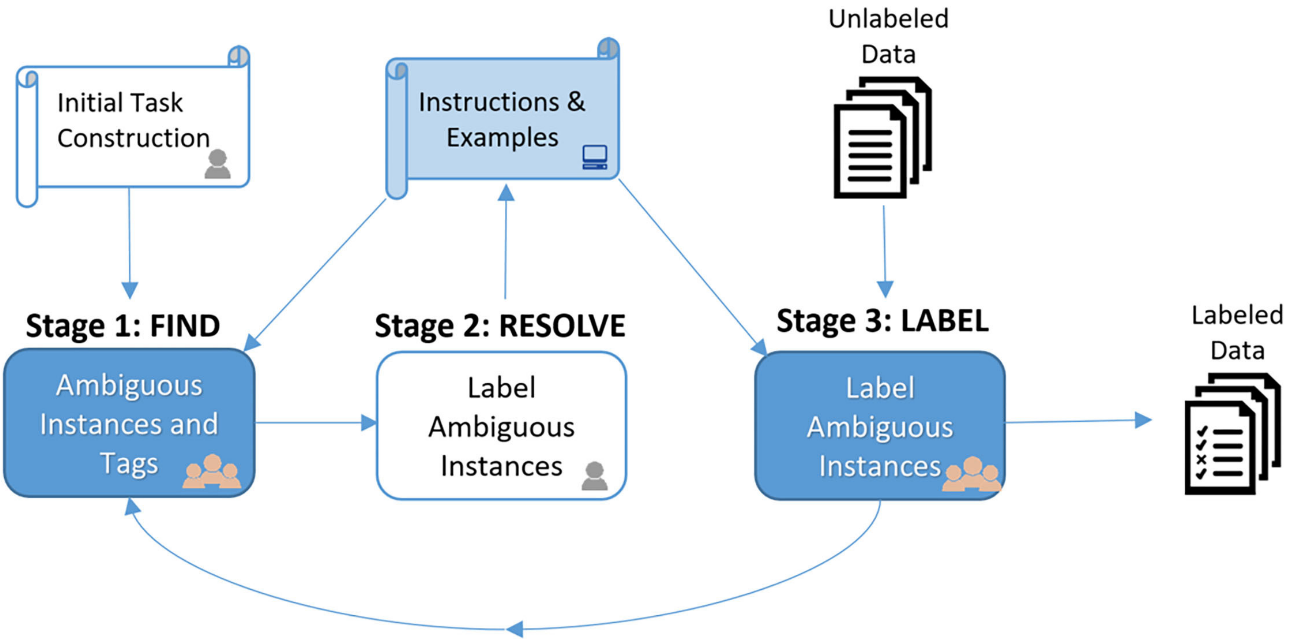

Initial Task Construction
Instructions & Examples
Unlabeled Data
Stage 1: FIND
Ambiguous Instances and Tags
Stage 2: RESOLVE
Label Ambiguous Instances
Stage 3: LABEL
Label Ambiguous Instances
Labeled Data

FIGURE 1 | Our Three-Stage FIND-RESOLVE-LABEL workflow is shown above. Stage 1 (FIND) asks the crowd to find examples whose correct label seems ambiguous given the task instructions (e.g., using external Internet search or database lookup). In Stage 2 (RESOLVE), the requester selects and labels one or more of these ambiguous examples. These are then automatically injected back into task instructions in order to improve clarity. Finally, in Stage 3 (LABEL), workers perform the actual annotation using the revised guidelines with clarifying examples. If Stage 3 labeling quality is insufficient, we can return to Stage 1 to find more ambiguous examples to further clarify instructions.

hand, also bring with them their own variety of implicit biases which the requester may not detect or understand (Ipeirotis et al., 2010; Sen et al., 2015; Dumitrache et al., 2018; Geva et al., 2019; Al Kuwatly et al., 2020; Fazelpour and De-Arteaga, 2022).

2.1. Helping Requesters Succeed

2.1.1. Best Practices

A variety of tutorials, surveys, introductions, and research articles offer how-to advice for successful microtask crowdsourcing with platforms such as MTurk (Jones, 2013; Marshall and Shipman, 2013; Egelman et al., 2014; Kovashka et al., 2016). For example, it is often recommended that requesters invest time browsing and labeling some data themselves before launching a task in order to better define and debug it (Alonso, 2015). Studies have compared alternative task designs to suggest best practices (Grady and Lease, 2010; Kazai et al., 2011; Papoutsaki et al., 2015; Wu and Quinn, 2017).

2.1.2. Templates and Assisted Design

Rather than start task design from scratch, MTurk offers templates and has suggested that requesters share successful templates for others' use (Chen et al., 2011). Similarly, classic research on software *design patterns* (Gamma et al., 1995) has inspired ideas for similar crowdsourcing design patterns which could be reused across different data collection tasks. For example, FIND-FIX-VERIFY (Bernstein et al., 2010) is a well-known example that partially inspired our study. Other researchers have suggested improved tool support for workflow design (Kittur et al., 2012) or engaging the crowd itself in task design or decomposition (Kittur et al., 2011; Kulkarni et al., 2012a).

2.1.3. Automating Task Design

Other researchers have gone further still to propose new middleware and programmable APIs to let requesters define tasks more abstractly and leave some design and management tasks to the middleware (Little et al., 2010b; Ahmad et al., 2011; Franklin et al., 2011; Barowy et al., 2016; Chen et al., 2016).

2.2. Understanding Disagreement

2.2.1. Random Noise vs. Bias

Since annotators are human, even trusted annotators will naturally make mistakes from time to time. Fortunately, random error is exactly the kind of disagreement that aggregation (Hung et al., 2013; Sheshadri and Lease, 2013) can easily resolve; assuming such mistakes are relatively infrequent and independent, workers will rarely err at the same instance, and therefore, techniques as simple as majority voting can address random noise. On the other hand, if workers have individual biases, they will make consistent errors; e.g., a teenager vs. a protective parent might have liberal vs. conservative biases in rating movies (Ipeirotis et al., 2010). In this case, it is useful to detect such consistent biases and re-calibrate worker responses to undo such biases. Aggregation can also work provided that workers do not share the same biases. However, when workers do share systematic biases, the independence assumption underlying aggregation is violated, and so aggregation can amplify bias rather than resolve it. Consequently, it is important that task design annotation guidelines should be vetted to ensure they identify cases in which annotator biases conflict with desired labels and particularly establish clear expectations for how such cases should be handled (Draws et al., 2021; Nouri et al., 2021b).

2.2.2. Objective vs. Subjective Tasks

In fully-objective tasks, we assume each question has a single correct answer, and any disagreement with the gold standard reflects error. Label aggregation methods largely operate in this space. On the other extreme, purely-subjective (i.e., opinion) tasks permit a wide range of valid responses with little expectation of agreement between individuals (e.g., asking about one's favorite color or food). Between these simple extremes, however, lies a wide, interesting, and important space of partially-subjective tasks in which answers are only partially-constrained (Tian and Zhu, 2012; Sen et al., 2015; Nguyen et al., 2016). For example, consider rating item quality: while agreement tends to be high for items having extremely good or bad properties, instances with more middling properties naturally elicit a wider variance in opinion. In general, because subjectivity permits a valid diversity of responses, it can be difficult to detect if an annotator does not undertake a task in good faith, complicating quality assurance.

2.2.3. Difficulty vs. Ambiguity

Some annotation tasks are more complex than others, just as some instances within each task are more difficult to label than other instances. A common concern with crowdsourcing is whether inexpert workers have sufficient expertise to successfully undertake a given annotation task. Intuitively, more guidance and scaffolding are likely necessary with more skilled tasks and fewer expert workers (Huang et al., 2021). Alternatively, if we use sufficiently expert annotators, we assume difficult cases can be handled (Retelny et al., 2014; Vakharia and Lease, 2015). With ambiguity, on the other hand, it would be unclear even to an expert what to do. Ambiguity is an interaction between data instances and annotation guidelines; effectively, an ambiguous instance is a corner-case with respect to guidelines. Aggregation can helpfully identify the majority interpretation but that interpretation may or may not be what is actually desired. Both difficult and ambiguous cases can lead to label confusion. Krivosheev et al. (2020) developed mechanisms to efficiently detect label confusion in classification tasks and demonstrated that alerting workers of the risk of confusion can improve annotation performance.

2.2.4. Static vs. Dynamic Disagreement

As annotators undertake a task, their understanding of work evolves as they develop familiarity with both the data and the guidelines. In fact, prior study has shown that annotators interpret and implement task guidelines in different ways as annotation progresses (Scholer et al., 2013; Kalra et al., 2017). Consequently, different sorts of disagreement can occur at different stages of annotation. Temporally-aware aggregation can partially ameliorate this (Jung and Lease, 2015), as can implementing data collection processes to train, "burn-in," or calibrate annotators, controlling, and/or accelerating their transition from an initial learning state into a steady state (Scholer et al., 2013). For example, we emphasize identifying key boundary cases and expected labels for them.

2.3. Mitigating Imperfect Instructions

Unclear, confusing, and ambiguous task instructions are commonplace phenomena on crowdsourcing platforms (Gadiraju et al., 2017; Wu and Quinn, 2017). In early study, Alonso et al. (2008) recommended collecting optional, free-form, task-level feedback from workers. While Alonso et al. (2008) found that some workers did provide example-specific feedback, the free-form nature of their feedback request elicited a variety of response types, which is difficult to check or to invalidate spurious responses. Alonso et al. (2008) also found that requiring such feedback led many workers to submit unhelpful text that was difficult to automatically cull. Such feedback was, therefore, recommended to be kept entirely optional.

While crowd work is traditionally completed independently to prevent collusion and enable statistical aggregation of uncorrelated work (Hung et al., 2013; Sheshadri and Lease, 2013; Zheng et al., 2017), a variety of work has explored collaboration mechanisms by which workers might usefully help each other complete a task more effectively (Dow et al., 2012; Kulkarni et al., 2012b; Drapeau et al., 2016; Chang et al., 2017; Manam and Quinn, 2018; Schaekermann et al., 2018; Chen et al., 2019; Manam et al., 2019).

Drapeau et al. (2016) proposed an asynchronous two-stage *Justify-Reconsider* method. In the Justify task, workers provide a rationale along with their answer referring to the task guidelines taught during training. For the Reconsider task, workers are confronted with an argument for the opposing answer submitted by another worker and then asked to reconsider (i.e., confirm or change) their original answer. The authors report that their Justify-Reconsider method generally yields higher accuracy but that requesting justifications requires additional cost. Consequently, they find that simply collecting more crowd annotations yields higher accuracy in a fixed-budget setting.

Chang et al. (2017) proposed a three-step approach in which crowd workers label the data, provide justifications for cases in which they disagree with others, and then review others' explanations. They evaluate their method on an image labeling task and report that requesting only justifications (without any further processing) does not increase the crowd accuracy. Their open-ended text responses can be subjective and difficult to check.

Kulkarni et al. (2012b) provide workers with a chat feature that supports workers in dealing with inadequate task explanations, suggesting additional examples to be given to requesters, teaching other workers how to use the UI, and verifying their hypotheses of the underlying task intent. Schaekermann et al. (2018) investigate the impact of discussion among crowd workers on the label quality using a chat platform allowing synchronous group discussion. While the chat platform allows workers to better express their justification than text excerpts, the discussion increases task completion times. In addition, chatting does not impose any restriction on the topic, limiting discussion from unenthusiastic workers and efficacy. Chen et al. (2019) also proposed a workflow allowing simultaneous discussion among crowd workers, and designed task instructions and a training phase to achieve effective discussions. While their method yields high labeling accuracy, the increased cost due to the discussion limits its task scope. Manam and Quinn (2018) evaluated both asynchronous and synchronous Q&A between workers and requesters to allow workers to ask questions to resolve any uncertainty about overall task instructions or specific examples. Bragg and Weld (2018) proposed an iterative workflow in which data instances with the low inter-rater agreement are put aside and either used as difficult training examples (if considered resolvable with respect to the current annotation guidelines) or used to refine the current annotation guidelines (if considered ambiguous).

Other study has explored approaches to address ambiguities even *before* the annotation process commences. For example, Manam et al. (2019) proposed a multi-step workflow enlisting the help of crowd workers to identify and resolve ambiguities in textual instructions. Gadiraju et al. (2017) and Nouri et al. (2021a) both developed predictive models to automatically score textual instructions for their overall level of clarity and Nouri et al. (2021b) proposed an interactive prototype to surface the predicted clarity scores to requesters in real-time as they draft and iterate on the instructions. Our approach also aims to resolve ambiguities upfront but focuses on identifying concrete visual examples of ambiguity and automatically enriching the underlying set of textual instructions with those examples.

Ambiguity arises from the interaction between annotation guidelines and particular data instances. Searching for ambiguous data instances within large-scale datasets or even the Internet can amount to finding a needle in a haystack. There exists an analogous problem of identifying "unknown unknowns" or "blind spots" of machine learning models. Prior study has proposed crowdsourced or hybrid human-machine approaches for spotting and mitigating model blind spots (Attenberg et al., 2011; Vandenhof, 2019; Liu et al., 2020). Our study draws inspiration from these workflows. We leverage the scale, intelligence, and common sense of the crowd to identify potential ambiguities within annotation guidelines and may, thus, aid in the process of mitigating blind spots in downstream model development.

2.4. Crowdsourcing Beyond Data Labeling

While data labeling represents the most common use of crowdsourcing in regard to training and evaluating machine learning models, human intelligence can be tapped in a much wider and more creative variety of ways. For example, the crowd might verify output from machine learning models, identify, and categorize blind spots (Attenberg et al., 2011; Vandenhof, 2019) and other failure modes (Cabrera et al., 2021), and suggest useful features for a machine learning classifier (Cheng and Bernstein, 2015).

One of the oldest crowdsourcing design patterns is utilizing the scale of the crowd for efficient, distributed exploration or filtering of large search spaces. Classic examples include the search for extraterrestrial intelligence[1], for Jim Gray's sailboat (Vogels, 2007) or other missing people (Wang et al., 2010), for DARPA's red balloons (Pickard et al., 2011), for astronomical events of interest (Lintott et al., 2008), and for endangered wildlife (Rosser and Wiggins, 2019) or bird species (Kelling et al., 2013). Across such examples, what is being sought must be broadly recognizable so that the crowd can accomplish the search task without the need for subject matter expertise (Kinney et al., 2008). In the 3-stage FIND-FIX-VERIFY crowdsourcing workflow (Bernstein et al., 2010), the initial FIND stage directs the crowd to identify "patches" in an initial text draft where more work is needed.

Our asking the crowd to search for ambiguous examples given task guidelines further explores the potential of this same crowd design pattern for distributed search. Rather than waiting for ambiguous examples to be encountered by chance during the annotation process, we instead seek to rapidly identify corner-cases by explicitly searching for them. We offload to the crowd the task of searching for ambiguous cases, and who better to identify potentially ambiguous examples than the same workforce that will be asked to perform the actual annotation? At the same time, we reduce requester work, limiting their effort to labeling corner-cases rather than adjusting the textual guidelines.

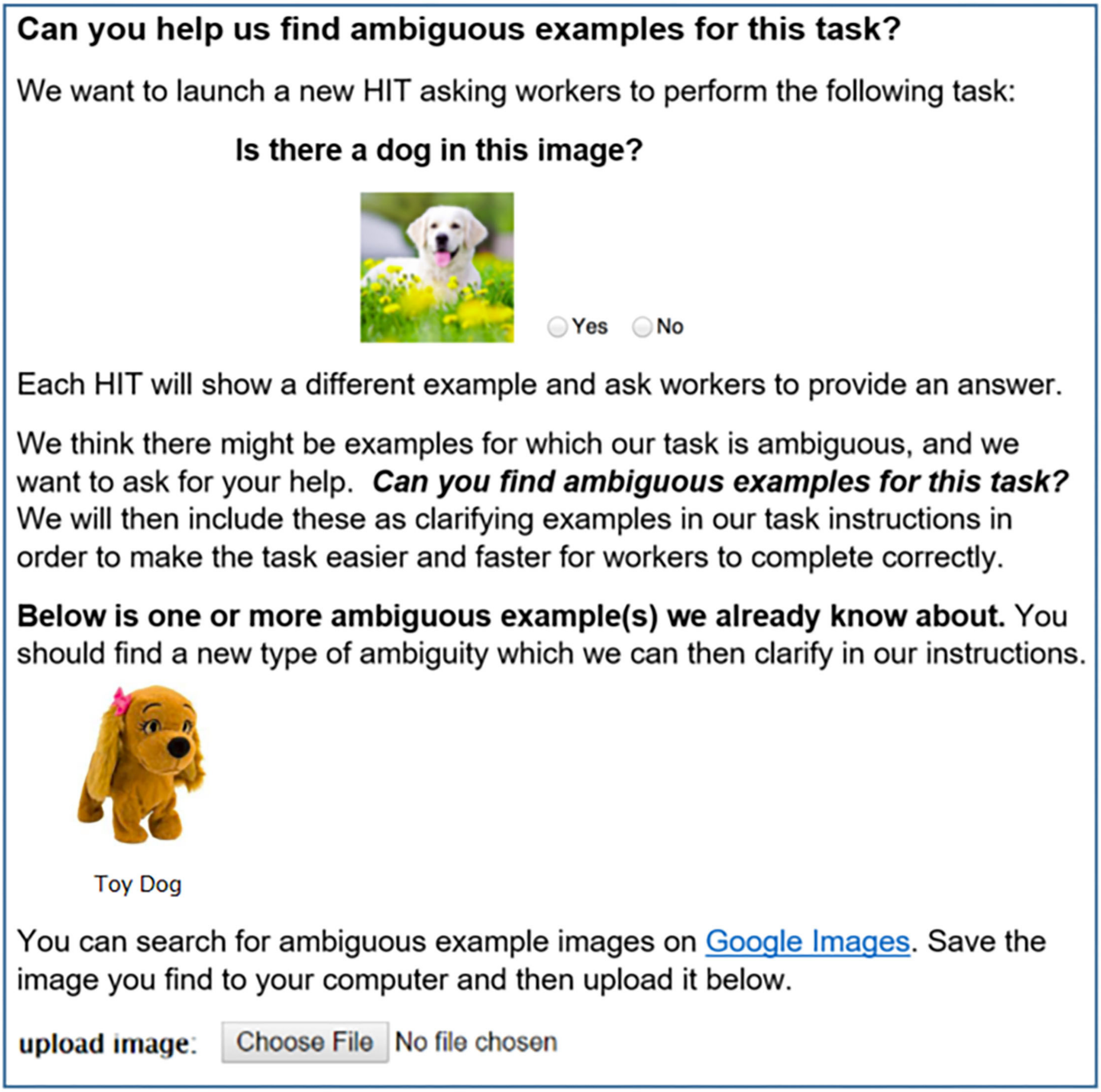
3. WORKFLOW DESIGN

In this study, we propose a three-stage FIND-RESOLVE-LABEL workflow for clarifying ambiguous corner cases in task instructions, investigated in the specific context of a binary image labeling task. An illustration of the workflow is shown in **Figure 1**. In Stage 1 (FIND), workers are asked to proactively collect ambiguous examples and concept tags given task instructions (section 3.1). Next, in Stage 2 (RESOLVE), the requester selects and labels one or more of the ambiguous examples found by the crowd. These labeled examples are then automatically injected back into task instructions in order to improve clarity (section 3.2). Finally, in Stage 3 (LABEL), workers perform the actual annotation task using the revised guidelines with clarifying examples (section 3.3). Requesters run the final LABEL stage on a sample of data, assess label quality, and then decide how to proceed. If quality is sufficient the remaining data can be labeled according to the current revision of the guidelines. Otherwise, Stages 1 and 2 can be repeated in order to further refine the clarity of annotation guidelines.

3.1. Stage 1: Finding Ambiguous Examples

In Stage 1 (FIND), workers are asked to collect ambiguous examples given the task instructions. For each ambiguous example, workers are also asked to generate a concept tag. The concept tag serves multiple purposes. First, it acts as a rationale (McDonnell et al., 2016; Kutlu et al., 2020), requiring workers to justify their answers and thus nudging them toward high-quality selections. Rationales also provide a form of transparency to help requesters better understand worker intent. Second, the concept tag provides a conceptual explanation of the ambiguity which

[1] https://setiathome.berkeley.edu/.



Can you help us find ambiguous examples for this task?

We want to launch a new HIT asking workers to perform the following task:

Is there a dog in this image?

○ Yes ○ No

Each HIT will show a different example and ask workers to provide an answer.

We think there might be examples for which our task is ambiguous, and we want to ask for your help. ***Can you find ambiguous examples for this task?*** We will then include these as clarifying examples in our task instructions in order to make the task easier and faster for workers to complete correctly.

Below is one or more ambiguous example(s) we already know about. You should find a new type of ambiguity which we can then clarify in our instructions.

Toy Dog

You can search for ambiguous example images on Google Images. Save the image you find to your computer and then upload it below.

upload image: Choose File No file chosen

FIGURE 2 | In the Stage 1 (FIND) task, workers are asked to search for examples they think would be ambiguous given task instructions. In this case, "Is there a dog in this image?" In collaboration conditions (section 3.1.1), workers will see additional ambiguous examples found by past workers.

can then be re-injected into annotation guidelines to help explain corner cases to future workers.

Figure 2 shows the main task interface for Stage 1 (FIND). The interface presents the annotation task (e.g., *"Is there a dog in this image?"*) and asks workers: *"Can you find ambiguous examples for this task?"* Pilot experiments revealed that workers had difficulty understanding the task based on this textual prompt alone. We, therefore, make the additional assumption that requesters provide a single ambiguous example to clarify the FIND task for workers. For example, the FIND stage for a dog annotation task could show the image of a Toy Dog as an ambiguous seed example. Workers are then directed to use Google Image Search to find these ambiguous examples. Once an ambiguous image is uploaded, another page (not shown) asks workers to provide a short concept tag summarizing the type of ambiguity represented by the example (e.g., *Toy Dog*).

3.1.1. Exploring Collaboration

To investigate the potential value of worker collaboration in finding higher quality ambiguities, we explore a light-weight, iterative design in which workers do not directly interact with each other, but are shown examples found by past workers (in addition to the seed example provided by the requester). For example, worker 2 would see an example selected by worker 1, and worker 3 would see examples found by workers 1 and 2, etc. Our study compares three different collaboration conditions described in section 4.4.1 below.

3.2. Stage 2: Resolving Ambiguous Examples

After collecting ambiguous examples in Stage 1 (FIND), the requester then selects and labels one or more of these examples. The requester interface for Stage 2 (RESOLVE) is shown in **Figure 3**. Our interface design affords a low-effort interaction in which requesters toggle examples between three states *via* mouse click: (1) selected as a positive example, (2) selected as a negative example, (3) unselected. Examples are unselected by default. The selected (and labeled) examples are injected back into the task instructions for Stage 3 (LABEL).

3.3. Stage 3: Labeling With Clarifying Examples

Best practices suggest that along with task instructions, requesters should include a set of examples and their correct annotations (Wu and Quinn, 2017). We automatically append to task instructions the ambiguous examples selected by the requester in Stage 2 (RESOLVE), along with their clarifying labels (**Figure 4**). Positive examples are shown first (*"you should select concepts like these"*), followed by negative examples (*"and NOT select concepts like these"*). Note that this stage does not require additional effort (e.g., instruction drafting) from the side of the requester because it merely augments the pre-existing task instruction template with the resulting list of clarifying examples.

4. METHODS

Experiments were conducted in the context of binary image classification. In particular, we designed six annotation tasks representing different variations of labeling for the presence or absence of *dog*-related concepts. Similar to prior study by Kulesza et al. (2014), we found this seemingly simplistic domain effective for our study because non-expert workers bring prior intuition as to how the classification could be done, but the task is characterized by a variety of subtle nuances and inherent ambiguities. We employed a between-subjects design in which each participant was assigned to exactly one experimental condition to avoid potential learning effects. This design was enforced using "negative" qualifications (Amazon Mechanical Turk, 2017) preventing crowd workers from participating in more than a single task. For the purpose of experimentation, authors acted as requesters. This included the specification of task instructions and intents and performing Stage 2 (RESOLVE), i.e., selecting clarifying examples for use in Stage 3 (LABEL).

4.1. Participant Recruitment and Quality Control

We recruited participants on Amazon's Mechanical Turk using workers from the US who had completed at least 1,000 tasks with a 95% acceptance rate. This filter served as a basic quality assurance mechanism to increase the likelihood of recruiting good-faith workers over "spammers." No further quality control mechanism was employed in our study to emulate imperfect, yet commonplace crowdsourcing practices for settings where definitive gold standard examples are not readily available for quality assessment. For ecological validity, we opted to not collect demographic information about participants prior to the annotation tasks.

4.2. Dataset

All experiments utilized the same set of 40 images. The image set was designed to encompass both easy, unambiguous cases and a range of difficult, ambiguous cases with respect to the question *"Is there a dog in this image?"* We first assembled a set of candidate images using a combination of (1) an online image search conducted by the authors to identify a set of clear positive and clear negative examples, (2) the Stage 1 (FIND) mechanism in which crowd workers on Amazon's Mechanical Turk identified difficult, ambiguous cases. Similar to Kulesza et al. (2014), we identified a set of underlying, dog-related categories *via* multiple passes of structured labeling on the data. From this process, 11 categories of dog-related concepts emerged: (1) dogs, (2) small dog breeds, (3) similar animals (easy to confuse with dogs), (4) cartoons, (5) stuffed toys, (6) robots, (7) statues, (8) dog-related objects (e.g., dog-shaped cloud), (9) miscellaneous (e.g., hot dog, the word "dog"), (10) different animals (difficult to confuse with dogs), and (11) planes (the easiest category workers should never confuse with dogs). Each image was assigned to exactly one category.

4.3. Annotation Tasks

When users of a search engine type in the query "apple," are they looking for information about the fruit, the company, or something else entirely? Despite the paucity of detail provided by a typical terse query, search result accuracy is assessed based on how well results match the user's underlying intent. Similarly, requesters on crowdsourcing platforms expect workers to understand the annotation "intent" underlying the explicit instructions provided. Analogously, worker accuracy is typically evaluated with respect to how well-annotations match that requester's intent even if instructions are incomplete, unclear, or ambiguous.

To represent this common scenario, we designed three different annotation tasks. For each task, the textual instructions exhibit a certain degree of ambiguity such that adding clarifying examples to instructions can help clarify requester intent to workers.

For each of the three tasks, we also selected two different intents, one more intuitive than the other in order to assess the effectiveness of our workflow design under intents of varying intuitiveness. In other words, we intentionally included one slightly more esoteric intent for each task hypothesizing that

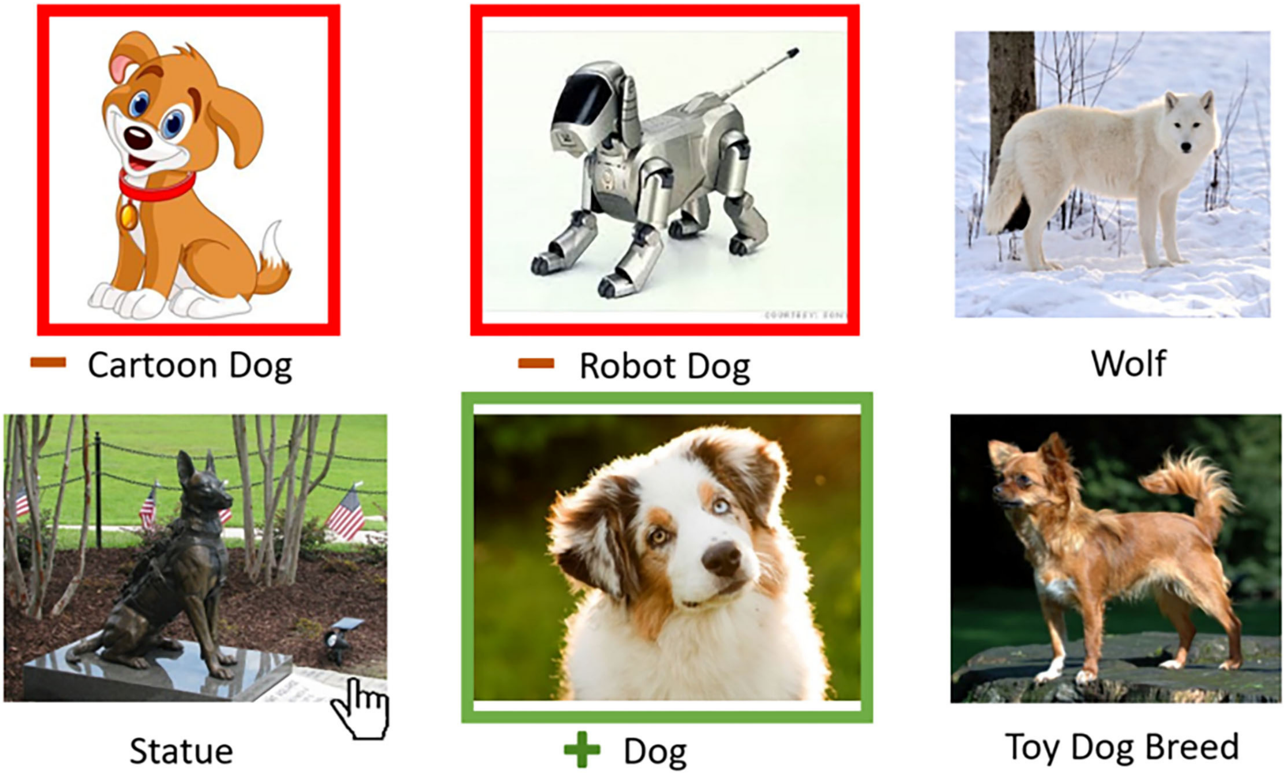


FIGURE 3 | For Stage 2 (RESOLVE), our interface design lets a requester easily select and label images. Each mouse click on an example toggles between unselected, selected positive, and selected negative states.

these would require workers to adapt to classification rules in conflict with their initial assumptions about requester intent. For each intent below, we list the categories constituting the positive class. All other categories are part of the negative class for the given intent.

For each of our six binary annotation tasks below, we partitioned examples into positive vs. negative classes given the categories included in the intent. We then measured worker accuracy in correctly labeling images according to positive and negative categories for each task.

4.3.1. Task 1: Is There a Dog in This Image?

Intent *a* (more intuitive): dogs, small dog breeds

Intent *b* (less intuitive): dogs, small dog breeds, similar animals. *Scenario*: The requester intends to train a machine learning model for avoiding animals and believes the model may also benefit from detecting images of wolves and foxes.

4.3.2. Task 2: Is There a Fake Dog in This Image?

Intent *a* (more intuitive): similar animals. *Scenario*: The requester is looking for animals often confused with dogs.

Intent *b* (less intuitive): cartoons, stuffed toys, robots, statues, objects. *Scenario*: The requester is looking for inanimate objects representing dogs.

4.3.3. Task 3: Is There a Toy Dog in This Image?

Intent *a* (less intuitive): small dog breeds. *Scenario*: Small dogs, such as Chihuahua or Yorkshire Terrier, are collectively referred to as "toy dog" breeds[2]. However, this terminology is not necessarily common knowledge making this intent less intuitive.

Intent *b* (more intuitive): stuffed toys, robots. *Scenario*: The requester is looking for children's toys, e.g., to train a model for an e-commerce site.

4.4. Evaluation

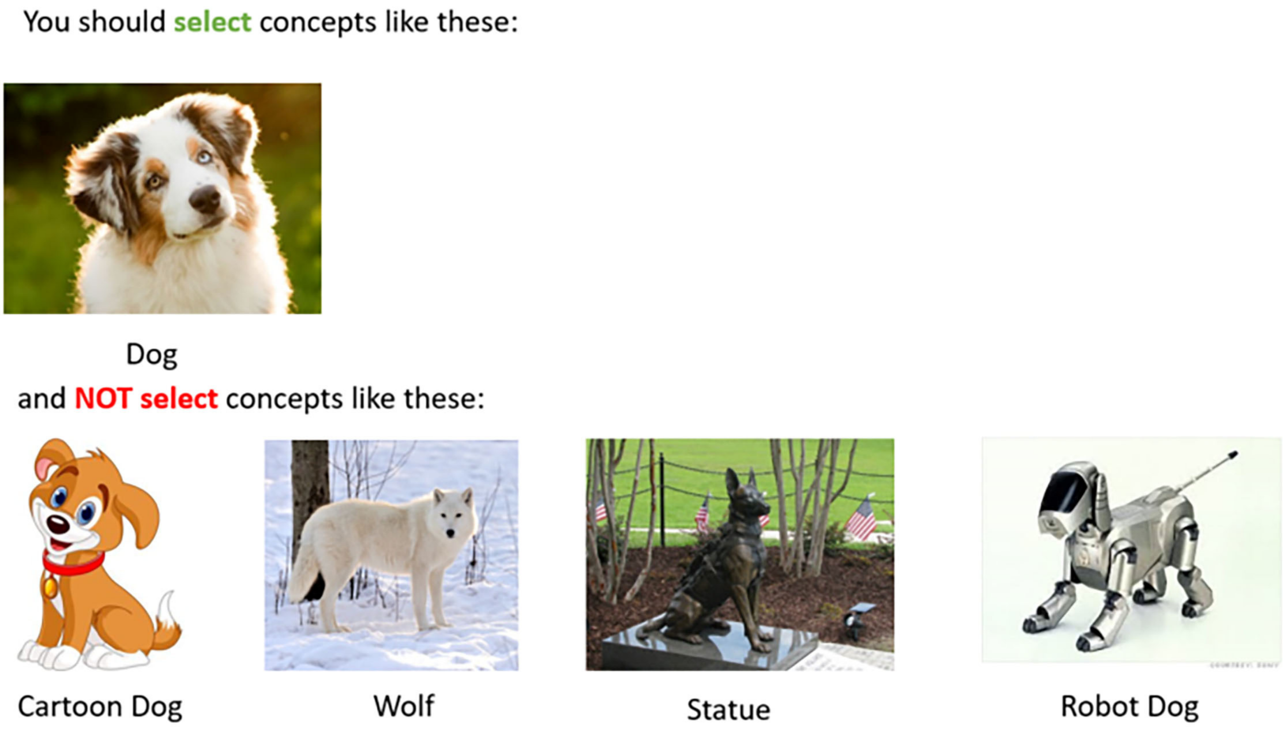
4.4.1. Qualitative Evaluation of Ambiguous Examples From Stage 1 (FIND)

For Stage 1 (FIND), we evaluated crowd workers' ability to find ambiguous images and concept tags for Task 1: *"Is there a dog in this image?"*. Through qualitative coding, we analyzed worker submissions based on three criteria: (1) correctness; (2) uniqueness; and (3) usefulness.

Correctness captures our assessment of whether the worker appeared to have understood the task correctly and submitted a plausible example of ambiguity for the given task. Any incorrect examples were excluded from consideration for uniqueness or usefulness.

Uniqueness captures our assessment of how many distinct types of ambiguity workers found across correct examples. For

[2] https://en.wikipedia.org/wiki/Toy_dog.



You should select concepts like these:

and NOT select concepts like these:

FIGURE 4 | For Stage 3 (LABEL), we combine the ambiguous instances and/or tags collected in Stage 1 (FIND) with the requester labels from Stage 2 (RESOLVE) and automatically inject the labeled examples back into task instructions.

example, we deemed "Stuffed Dog" and "Toy Dog" sufficiently close as to represent the same concept.

Usefulness captures our assessment of which of the unique ambiguous concepts found were likely to be useful in the annotation. For example, while an image of a hot dog is valid and unique, it is unlikely that many annotators would find it ambiguous in practice.

Our study compares two different collaboration conditions for Stage 1 (FIND). In both conditions, workers were shown one or more ambiguous examples with associated concept tags and were asked to add another, different example of ambiguity, along with a concept tag for that new example:

1. **No collaboration**. Each worker sees the task interface seeded with a single ambiguous example and its associated concept tag provided by the requester. Workers find additional ambiguous examples *independently* from other workers.
2. **Collaboration**. Workers see all ambiguous examples and their concept tags previously found by other workers. There is no filtering mechanism involved, so workers may be presented with incorrect and/or duplicated examples. This workflow configuration amounts to a form of unidirectional, asynchronous communication among workers.

For both collaboration conditions, a total of 15 ambiguous examples (from 15 unique workers) were collected and evaluated with respect to the above criteria.

4.4.2. Quantitative Evaluation of Example Effectiveness in Stage 3 (LABEL)

To evaluate the effectiveness of enriching textual instructions with ambiguous examples from Stage 1 (FIND) and to assess the relative utility of presenting workers with images and/or concept tags from ambiguous examples, we compared the following five conditions. The conditions varied in how annotation instructions were presented to workers in Stage 3 (LABEL):

1. **B0**: No examples were provided along with textual instructions.
2. **B1**: A set of randomly chosen examples were provided along with textual instructions.
3. **IMG**: Only images (but no concept tags) of ambiguous examples were shown to workers along with textual instructions.
4. **TAG**: Only concept tags (but no images) of ambiguous examples were shown to workers along with textual instructions.
5. **IMG+TAG**: Both images and concept tags of ambiguous examples were shown to workers along with textual instructions.

Each of the five conditions above was completed by nine unique workers. Each task consisted of classifying 10 images. Workers were asked to classify each of the 10 images into either the positive or the negative class.

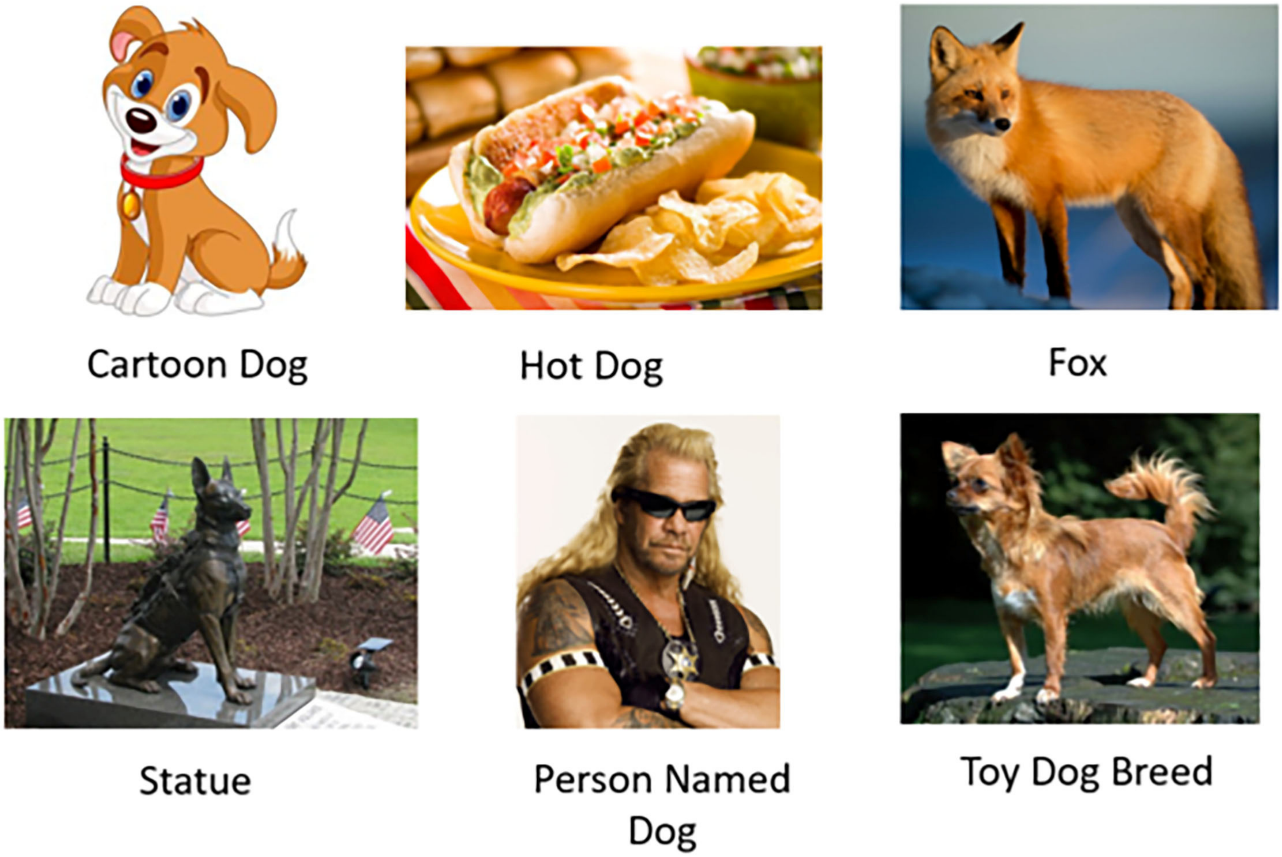


FIGURE 5 | Ambiguous examples and concept tags provided by workers in Stage 1 (FIND) for the task "Is there a dog in this image?" We capitalize tags here for presentation but use raw worker tags without modification in our evaluation.

5. RESULTS

5.1. Can Workers Find Ambiguous Concepts?

In this section, we provide insights from pilots of Stage 1 (FIND) followed by a qualitative analysis of ambiguous examples identified by workers in this stage.

5.1.1. Pilot Insights

5.1.1.1. Task Design

Initial pilots of Stage 1 (FIND) revealed two issues: (1) duplicate concepts, and (2) misunderstanding of the task. Some easy-to-find and closely related concepts were naturally repeated multiple times. One type of concept duplication was related to the *seed* example provided by requesters to clarify the task objective. In particular, some workers searched for additional examples of the *same* ambiguity rather than finding *distinct* instances of ambiguity. Another misunderstanding led some workers to submit *generally* ambiguous images, i.e., similar to Google Image Search results for search term "ambiguous image," rather than images that were ambiguous relative to the specific task instruction *"Is there a dog in this image?"* We acknowledge that our own task design was not immune to ambiguity, so we incorporated clarifications to instruct workers to find ambiguous examples *distinct* from the seed example and *specific* to the task instructions provided.

5.1.1.2. Unexpected Ambiguous Concepts

However, our pilots also revealed workers' ability to identify surprising examples of ambiguous concepts we had not anticipated. Some of these examples were educational and helped the paper authors learn about the nuances of our task. For example, one worker returned an image of a Chihuahua (a small dog breed) along with the concept tag "toy dog." In trying to understand the worker's intent, we learned that the term "toy dog" is a synonym for small dog breeds (see text footnote 2). Prior to that, our interpretation of the "toy dog" concept was limited to children's toys. This insight inspired Task 3 ("Is there a toy dog in this image?") with two different interpretations (section 4.3). Another unexpected ambiguous example was the picture of a man (**Figure 5**). We initially jumped to the conclusion that the worker's response was spam, but on closer inspection discovered that the picture displayed reality show celebrity "Dog the Bounty Hunter"[3] These instances are excellent illustrations of

[3] http://www.dogthebountyhunter.com.

TABLE 1 | Percentage of correct, unique, and useful examples from Stage 1 (FIND).

	Correct	Unique	Useful
No collaboration	60.0	26.7	26.7
Collaboration	**93.0**	**40.0**	**33.3**

The bold value indicates the largest values per column.

the possibility that crowd workers may interpret task instructions in valid and original ways entirely unanticipated by requesters.

5.1.2. Qualitative Assessment of Example Characteristics

We employed qualitative coding to assess whether worker submissions met each of the quality criteria (Correctness, Uniqueness, and Usefulness). **Table 1** shows the percentage of ambiguous examples meeting these criteria for the two conditions with and without collaboration, respectively. Our hypothesis that collaboration among workers can help produce higher quality ambiguous examples is supported by our results. Results show that, compared to no collaboration, a collaborative workflow produced substantially greater proportions of correct (93 vs. 60%), unique (40 vs. 27%), and useful (33 vs. 27%) ambiguous examples. A potential explanation for this result is that exposing workers to a variety of ambiguous concepts upfront may assist them in exploring the space of yet uncovered ambiguities more effectively.

5.2. Can Ambiguous Examples Improve Annotation Accuracy?

Next, we report quantitative results on how ambiguous examples—found in Stage 1 and selected and labeled in Stage 2—can be used as instructional material to improve annotation accuracy in Stage 3. We also provide an analysis of annotation errors.

5.2.1. Effectiveness of Ambiguous Examples

Our hypothesis is that these examples can be used to help delineate the boundary of our annotation task and, hence, teach annotation guidelines to crowd workers better than randomly chosen examples. **Table 2** reports crowd annotation accuracy for each of the six tasks broken down by experimental condition.

5.2.1.1. Using Examples to Teach Annotation Guidelines

Intuitively, providing examples to workers helps them to better understand the intended labeling task (Wu and Quinn, 2017). Comparing designs B0 and B1 in **Table 2**, we clearly see that providing examples (B1) almost always produces more accurate labeling than a design that provides no examples (B0). In addition to this, the IMG design performs better than B1. This shows that the kind of examples that are provided is also important. Showing ambiguous examples is clearly superior to showing randomly chosen examples. This supports our hypothesis: ambiguous examples appear to delineate labeling boundaries for the task better than random examples.

5.2.1.2. Instances vs. Concepts

Best practices suggest that requesters provide examples when designing their tasks (Wu and Quinn, 2017). We include this design in our evaluation as B1. An alternate design is to show concepts as examples instead of specific instances; this is our design TAG, shown in **Table 2**. For example, for the task "Is there a Dog in this image?", instead of showing a dog statue image, we could simply provide the example concept "Inanimate Objects" should be labeled as NO. Results in **Table 2** show that TAG consistently outperforms IMG, showing that teaching *via* example concepts can be superior to teach *via* example instances.

5.2.1.3. Concepts Only vs. Concepts and Examples

Surprisingly, workers who were presented shown only the concept tags performed better than workers who were shown concept tags along with an example image for each concept. Hence, the particular instance chosen may not represent the concept well. This might be overcome by better selecting a more representative example for a concept or showing more examples for each concept. We leave such questions for future study.

5.2.2. Sources of Worker Errors

5.2.2.1. Difficult vs. Subjective Questions

Table 3 shows accuracy for categories "Similar Animal" and "Cartoon" for Task 1b (section 4.3). We see that some concepts appear more difficult, such as correctly labeling a wolf or a fox. Annotators appear to need some world knowledge or training of differences between species in order to correctly distinguish such examples vs. dogs. Such concepts seem more difficult to teach; even though the accuracy improves, the improvement is less than we see with other concepts. In contrast, for Cartoon Dog (an example of a subjective question), adding this category to the illustrative examples greatly reduces the ambiguity for annotators. Other concepts like "Robot" and "Statue" also show large improvements in accuracy.

5.2.2.2. Learning Closely Related Concepts

To see if crowdworkers learn closely related concepts without being explicitly shown examples, consider "Robot Dog" and "Stuffed Toy" as two types of a larger "Toy Dog" children's toy concept. In Task 1b, the workers are shown the concept "Robot Dog" as examples labeled as NO, without being shown an example for "Stuffed Toy." **Table 3** shows that workers learn the related concept "Stuffed Toy" and accurately label the instances that belong to this concept. The performance gain for the concept "Toy Dog" is the same as the gain for "Robot Dog," when we compare design IMG+TAG and B1. Other similarly unseen concepts [marked with an asterisk (*) in the table] show that workers are able to learn the requester's intent for unseen concepts if given examples of other, similar concepts.

5.2.2.3. Peer Agreement With Ambiguous Examples

It is not always possible or cost-effective to obtain expert/gold labels for tasks, so requesters often rely on peer-agreement between workers to estimate worker reliability. Similarly, majority voting or weighted voting is often used to aggregate worker labels for consensus (Hung et al., 2013; Sheshadri and

TABLE 2 | Worker accuracy [%] for all six tasks by condition.

Design	Task 1a	Task 1b	Task 2a	Task 2b	Task 3a	Task 3b
B0	75.6	70.1	47.8	78.7	69.1	92.9
B1	83.0	66.4	59.0	85.5	78.7	96.0
IMG	88.0	89.5	68.5	85.8	89.2	93.5
TAG	91.0	**91.0**	79.0	**87.0**	**91.0**	**96.9**
IMG+TAG	**91.4**	87.0	**81.8**	86.4	88.3	96.6

The bold value indicates the largest values per column.

TABLE 3 | Worker accuracy [%] on Task 1b, broken down by concept category.

Design	Similar animal	Stuffed toy*	Robot	Statue	Cartoon	...
B0	37.0	22.2	33.3	18.5	62.2	
B1	14.8	22.2	25.9	29.6	57.8	
IMG	44.4	**100.0**	**100.0**	**92.6**	**100.0**	
TAG	**74.1**	**100.0**	88.9	88.9	97.8	
IMG+TAG	48.1	88.9	92.6	77.8	97.8	
Design	**Objects***	**Unseen ambiguity***	**Small breed**	**Plane**	**Dog**	**Another animal**
B0	74.1	**88.9**	88.9	**100.0**	**100.0**	**100.0**
B1	59.3	82.2	**94.4**	**100.0**	**100.0**	**100.0**
IMG	**100.0**	75.6	88.9	**100.0**	95.6	**100.0**
TAG	**100.0**	80.0	**94.4**	**100.0**	91.1	**100.0**
IMG+TAG	96.3	77.8	91.7	96.3	95.6	88.9

We see that some hard concepts cannot be easily disambiguated, e.g., Similar Animal. Concepts for which workers were not shown any examples are marked with an asterisk (*).
The bold value indicates the largest values per column.

TABLE 4 | Worker accuracy [%] on ambiguous vs unambiguous categories with baseline B1.

Task	Unambiguous	Ambiguous
1a	96.1	72.2
1b	98.6	**41.7**
2a	86.8	**42.2**
2b	98.5	82.5
3a	82.5	75.8
3b	98.6	93.8

For Tasks 1b and 2a, the majority vote (among nine workers) on ambiguous examples is wrong.
The bold values indicate substantially lower than the other values in the table.

Lease, 2013). However, we also know that when workers have consistent, systematic group biases, the aggregation will serve to reinforce and amplify the group bias rather than mitigate it (Ipeirotis et al., 2010; Sen et al., 2015; Dumitrache et al., 2018; Fazelpour and De-Arteaga, 2022).

While we find agreement often correlates with accuracy, and so have largely omitted to report it in this study, we do find several concepts for which the majority chooses wrong answers, producing high agreement but low accuracy. Recall that our results are reported over nine workers per example, whereas typical studies use a plurality of three or five workers. Also recall that Tasks 1b and 2a (section 4.3) represent two of our less intuitive annotation tasks for which requester intent may be at odds with worker intuition, requiring greater task clarity for over-coming worker bias.

Table 4 shows majority vote accuracy for these tasks for the baseline B1 design which (perhaps typical of many requesters) includes illustrative examples but not necessarily the most informative ones. Despite collecting labels from nine different workers, the majority is still wrong, with majority vote accuracy on ambiguous examples falling below 50%.

6. CONCLUSION AND FUTURE WORK

6.1. Summary and Contributions

Quality assurance for labeled data remains a challenge today. In chasing the potential advantages crowdsourcing has to offer, some important quality assurance best practices from traditional export annotation workflows may be lost. Our work adds to the existing literature on mechanisms to disambiguate nuanced class categories and, thus, improve labeling guidelines and, in effect, classification decisions in crowdsourced data annotation.

In this study, we presented a three-stage FIND-RESOLVE-LABEL workflow as a novel mapping of traditional annotation processes, involving iterative refinement of guidelines by expert annotators, onto a light-weight, structured task design suitable for crowdsourcing. Through careful task design and intelligent distribution of effort between crowd workers and requesters, it may be possible for the crowd to play a valuable role in reducing requester effort while also helping requesters to better understand the nuances and edge cases of their intended annotation taxonomy in order to generate clearer task instructions for the crowd. In contrast to prior work, our approach is proactive and open-ended, leveraging crowd workers' unconstrained creativity and intelligence to identify ambiguous examples online through an Internet search, proactively enriching task instructions with these examples upfront before the annotation process commences.

While including illustrative examples in instructions is known to be helpful (Wu and Quinn, 2017), we have shown that not all examples are equally informative to annotators and that intelligently selecting ambiguous corner-cases can improve labeling quality. Our results revealed that the crowd performed worst on ambiguous instances and, thus, can benefit the most from help for cases where requester intents run counter to annotators' internal biases or intuitions. For some instances of ambiguity, we observed high agreement among workers on answers contrary to what the requester defines as correct. Such tasks are likely to produce an incorrect label even when we employ intelligent answer aggregation techniques. Techniques like ours to refine instruction clarity are particularly critical in such cases.

Finally, we found that workers were able to infer the correct labels for concepts closely related to the target concept. This result suggests that it may not be necessary to identify and clarify *all* ambiguous concepts that could potentially be encountered during the task. An intelligently selected set of clarifying examples may enable the crowd to disambiguate labels of unseen examples accurately even if not all instances of ambiguity are exhaustively covered.

6.2. Limitations

In this study, we propose a novel workflow for addressing the issue of inherent ambiguity in data classification settings. However, our study is not without limitations. First, our study focuses on a specific type of annotation task (image classification). While our workflow design targets data classification tasks in general, further study is needed to empirically validate the usefulness of this approach for other data annotation settings.

Second, our approach is based on the assumption that characteristics of ambiguous instances contributed by the crowd *via* external Internet search will match those of the dataset being evaluated. However, this assumption may not always be met depending on the domain and modality of the dataset. Certain datasets may not be represented *via* publicly available external search. In that case, additional building blocks would be needed in the workflow to enable effective search over a private data repository. Existing solutions for external search may already cluster results based on representative groups or classes. Our empirical results leave open the question of to what extent this feature could have influenced or facilitated the task for workers.

Third, our evaluation is limited in terms of datasets, task types, and the size of our participant sample. Caution is warranted in generalizing our results beyond the specific evaluation setting, e.g., since characteristics of the dataset can influence the results. Given the limited size of our participant sample and the fact that crowd populations can be heterogeneous, our empirical data was amenable only to descriptive statistics but not to null hypothesis significance tests. In conclusion, our results should be considered indicative of the potential usefulness of our approach rather than being fully definitive. Further study is needed to validate our approach in a more statistically robust and generalizable manner *via* larger samples.

6.3. Future Study

While we evaluate our strategy on an image labeling task, our approach is more general and could be usefully extended to other domains and tasks. For example, consider collecting document relevance judgments in information retrieval (Alonso et al., 2008; Scholer et al., 2013; McDonnell et al., 2016), where user *information needs* are often subjective, vague, and incomplete. Such generalization may raise new challenges. For example, the image classification task used in our study allows us to point workers to online image searches. However, other domains may require additional or different search tools (e.g., access to collections of domain-specific text documents) for workers to be able to effectively identify ambiguous corner cases.

Alonso (2015) proposes having workers perform practice tasks to get familiarized with the data and thus increase annotation performance for future tasks. While our experimental setup prevented workers from performing more than one task to avoid potential learning effects, future study may explore and leverage workers' ability to improve their performance for certain types of ambiguity over time. For example, we may expect that workers who completed Stage 1 are better prepared for Stage 3 given that they have already engaged in the mental exercise of critically exploring the decision boundary of the class taxonomy in question.

Another best practice from LDC is deriving a decision tree for common ambiguous cases which annotators can follow as a principled and consistent way to determine the label of ambiguous examples (Griffitt and Strassel, 2016). How might we use the crowd to induce such a decision tree? Prior design study in effectively engaging the crowd in clustering (Chang et al., 2016) can guide design considerations for this challenge.

In our study, Stage 2 RESOLVE required requesters to select ambiguous examples. Future study may explore variants of Stage 1 FIND where requesters *filter* the ambiguous examples provided by the crowd. There is an opportunity for saving requester effort if both of these stages are combined. For instance, examples selected in the filtering step of Stage 1 can be fed forward to reduce the example set considered for labeling in Stage 2. Another method would be to have requesters perform labeling simultaneously with filtering in Stage 1, eliminating Stage 2 altogether. Finally, if the requester deems the label quality of Stage 3 insufficient and initiates another cycle of ambiguity reduction *via* Stages 1 and 2 those stages could start with examples already identified in the prior cycle.

A variety of other directions can be envisioned for further reducing requester effort. For example, the crowd could be called upon to verify and prune ambiguous examples collected in the initial FIND stage. Examples flagged as spam or assigned a low ambiguity rating could be automatically discarded to minimize requester involvement in Stage 2. Crowd ambiguity ratings could also be used to rank examples for guiding requesters' attention in Stage 2. A more ambitious direction for future study would be to systematically explore how well and under what circumstances the crowd is able to correctly infer requester intent. Generalizable insights about this question would enable researchers to design strategies that eliminate requester involvement altogether under certain conditions.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of an unrecoverable data loss. Data that is available will be shared online upon acceptance at https://www.ischool.utexas.edu/~ml/publications/. Requests to access the datasets should be directed to ML, ml@utexas.edu.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

VP: implementation, experimentation, data analysis, and manuscript preparation. MS: manuscript preparation. ML: advising on research design and implementation and manuscript preparation. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported in part by the Micron Foundation and by Good Systems (https://goodsystems.utexas.edu), a UT Austin Grand Challenge to develop responsible AI technologies.

ACKNOWLEDGMENTS

We thank our many talented crowd contributors, without whom our research would not be possible. We thank the reviewers for their time and assistance in helping us to further improve this study.

REFERENCES

Ahmad, S., Battle, A., Malkani, Z., and Kamvar, S. (2011). "The jabberwocky programming environment for structured social computing," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 53–64. doi: 10.1145/2047196.2047203

Al Kuwatly, H., Wich, M., and Groh, G. (2020). "Identifying and measuring annotator bias based on annotators' demographic characteristics," in *Proceedings of the Fourth Workshop on Online Abuse and Harms (at EMNLP)* (Association for Computational Linguistics), 184–190. doi: 10.18653/v1/2020.alw-1.21

Alonso, O. (2015). "Practical lessons for gathering quality labels at scale," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1089–1092. doi: 10.1145/2766462.2776778

Alonso, O., Rose, D. E., and Stewart, B. (2008). "Crowdsourcing for relevance evaluation," in *ACM SigIR Forum, Vol. 42*, 9–15. doi: 10.1145/1480506.1480508

Amazon Mechanical Turk (2017). *Tutorial: Best Practices for Managing Workers in Follow-Up Surveys or Longitudinal Studies*. Available online at: https://blog.mturk.com/tutorial-best-practices-for-managing-workers-in-follow-up-surveys-or-//longitudinal-studies-4d0732a7319b

Attenberg, J., Ipeirotis, P. G., and Provost, F. (2011). "Beat the machine: challenging workers to find the unknown unknowns," in *Proceedings of the 11th AAAI Conference on Human Computation, AAAIWS'11-11* (AAAI Press), 2–7.

Barowy, D. W., Curtsinger, C., Berger, E. D., and McGregor, A. (2016). Automan: a platform for integrating human-based and digital computation. *Commun. ACM* 59, 102–109. doi: 10.1145/2927928

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., et al. (2010). "Soylent: a word processor with a crowd inside," in *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, 313–322. doi: 10.1145/1866029.1866078

Bragg, J., and Weld, D. S. (2018). "Sprout: crowd-powered task design for crowdsourcing," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 165–176. doi: 10.1145/3242587.3242598

Cabrera, A. A., Druck, A. J., Hong, J. I., and Perer, A. (2021). Discovering and validating ai errors with crowdsourced failure reports. *Proc. ACM Hum.-Comput. Interact.* 5: CSCW2. doi: 10.1145/3479569

Chang, J. C., Amershi, S., and Kamar, E. (2017). "Revolt: collaborative crowdsourcing for labeling machine learning datasets," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2334–2346. doi: 10.1145/3025453.3026044

Chang, J. C., Kittur, A., and Hahn, N. (2016). "Alloy: clustering with crowds and computation," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3180–3191. doi: 10.1145/2858036.2858411

Chen, J. J., Menezes, N. J., Bradley, A. D., and North, T. (2011). "Opportunities for crowdsourcing research on amazon mechanical Turk," in *ACM CHI Workshop on Crowdsourcing and Human Computation*.

Chen, Q., Bragg, J., Chilton, L. B., and Weld, D. S. (2019). "Cicero: multi-turn, contextual argumentation for accurate crowdsourcing," in *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*, 1–14. doi: 10.1145/3290605.3300761

Chen, Y., Ghosh, A., Kearns, M., Roughgarden, T., and Vaughan, J. W. (2016). Mathematical foundations for social computing. *Commun. ACM* 59, 102–108. doi: 10.1145/2960403

Cheng, J., and Bernstein, M. S. (2015). "Flock: hybrid crowd-machine learning classifiers," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 600–611. doi: 10.1145/2675133.2675214

Cole, C. (2011). A theory of information need for information retrieval that connects information to knowledge. *J. Assoc. Inform. Sci. Technol.* 62, 1216–1231. doi: 10.1002/asi.21541

Dow, S., Kulkarni, A., Klemmer, S., and Hartmann, B. (2012). "Shepherding the crowd yields better work," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1013–1022. doi: 10.1145/2145204.2145355

Drapeau, R., Chilton, L. B., Bragg, J., and Weld, D. S. (2016). "Microtalk: using argumentation to improve crowdsourcing accuracy," in *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Draws, T., Rieger, A., Inel, O., Gadiraju, U., and Tintarev, N. (2021). "A checklist to combat cognitive biases in crowdsourcing," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 9*, 48–59.

Dumitrache, A., Inel, O., Aroyo, L., Timmermans, B., and Welty, C. (2018). Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv [preprint] arXiv:1808.06080*. Available online at: https://arxiv.org/abs/1808.06080

Egelman, S., Chi, E. H., and Dow, S. (2014). "Crowdsourcing in HCI research," in *Ways of Knowing in HCI* (Springer), 267–289. doi: 10.1007/978-1-4939-0378-8_11

Fazelpour, S., and De-Arteaga, M. (2022). Diversity in sociotechnical machine learning systems. *arXiv [Preprint]*. arXiv: 2107.09163. Availalble online at: https://arxiv.org/pdf/2107.09163.pdf

Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., and Xin, R. (2011). "Crowddb: answering queries with crowdsourcing," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, 61–72. doi: 10.1145/1989323.1989331

Gadiraju, U., Yang, J., and Bozzon, A. (2017). "Clarity is a worthwhile quality: on the role of task clarity in microtask crowdsourcing," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 5–14. doi: 10.1145/3078714.3078715

Gaikwad, S. N. S., Whiting, M. E., Gamage, D., Mullings, C. A., Majeti, D., Goyal, S., et al. (2017). "The daemo crowdsourcing marketplace," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1–4. doi: 10.1145/3022198.3023270

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-oriented Software*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Geva, M., Goldberg, Y., and Berant, J. (2019). "Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong), 1161–1166. doi: 10.18653/v1/D19-1107

Goto, S., Ishida, T., and Lin, D. (2016). "Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 4*.

Grady, C., and Lease, M. (2010). "Crowdsourcing document relevance assessment with mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk* (Association for Computational Linguistics), 172–179.

Griffitt, K., and Strassel, S. (2016). "The query of everything: developing open-domain, natural-language queries for bolt information retrieval," in *LREC*.

Huang, G., Wu, M.-H., and Quinn, A. J. (2021). "Task design for crowdsourcing complex cognitive skills," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. doi: 10.1145/3411763.3443447

Hung, N. Q. V., Tam, N. T., Tran, L. N., and Aberer, K. (2013). "An evaluation of aggregation techniques in crowdsourcing," in *International Conference on Web Information Systems Engineering* (Springer), 1–15. doi: 10.1007/978-3-642-41154-0_1

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). "Quality management on amazon mechanical Turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 64–67. doi: 10.1145/1837885.1837906

Jones, G. J. (2013). "An introduction to crowdsourcing for language and multimedia technology research," in *Information Retrieval Meets Information Visualization* (Springer), 132–154. doi: 10.1007/978-3-642-36415-0_9

Jung, H. J., and Lease, M. (2015). "Modeling temporal crowd work quality with limited supervision," in *Proceedings of the 3rd AAAI Conference on Human Computation (HCOMP)*, 83–91.

Kalra, K., Patwardhan, M., and Karande, S. (2017). "Shifts in rating bias due to scale saturation," in *Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*.

Kazai, G., Kamps, J., Koolen, M., and Milic-Frayling, N. (2011). "Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 205–214. doi: 10.1145/2009916.2009947

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W.-K., Yu, J., et al. (2013). A human/computer learning network to improve biodiversity conservation and research. *AI Mag*. 34, 10. doi: 10.1609/aimag.v34i1.2431

Kinney, K. A., Huffman, S. B., and Zhai, J. (2008). "How evaluator domain expertise affects search result relevance judgments," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 591–598. doi: 10.1145/1458082.1458160

Kittur, A., Khamkar, S., André, P., and Kraut, R. (2012). "Crowdweaver: visually managing complex crowd work," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 1033–1036. doi: 10.1145/2145204.2145357

Kittur, A., Smus, B., Khamkar, S., and Kraut, R. E. (2011). "Crowdforge: crowdsourcing complex work," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 43–52. doi: 10.1145/2047196.2047202

Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K., et al. (2016). Crowdsourcing in computer vision. *Found. Trends Comput. Graph. Vis*. 10, 177–243. doi: 10.1561/0600000071

Krivosheev, E., Bykau, S., Casati, F., and Prabhakar, S. (2020). Detecting and preventing confused labels in crowdsourced data. *Proc. VLDB Endow*. 13, 2522–2535. doi: 10.14778/3407790.3407842

Kulesza, T., Amershi, S., Caruana, R., Fisher, D., and Charles, D. (2014). "Structured labeling for facilitating concept evolution in machine learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3075–3084. doi: 10.1145/2556288.2557238

Kulkarni, A., Gutheim, P., Narula, P., Rolnitzky, D., Parikh, T., and Hartmann, B. (2012b). Mobileworks: designing for quality in a managed crowdsourcing architecture. *IEEE Intern. Comput*. 16, 28–35. doi: 10.1109/MIC.2012.72

Kulkarni, A., Can, M., and Hartmann, B. (2012a). "Collaboratively crowdsourcing workflows with Turkomatic," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (IEEE), 1003–1012. doi: 10.1145/2145204.2145354

Kutlu, M., McDonnell, T., Elsayed, T., and Lease, M. (2020). Annotator rationales for labeling tasks in crowdsourcing. *J. Artif. Intell. Res*. 69, 143–189. doi: 10.1613/jair.1.12012

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the Sloan digital sky survey. *Monthly Not. R. Astron. Soc*. 389, 1179–1189. doi: 10.1111/j.1365-2966.2008.13689.x

Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010a). "Exploring iterative and parallel human computation processes," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 68–76. doi: 10.1145/1837885.1837907

Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010b). "Turkit: human computation algorithms on mechanical Turk," in *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, 57–66. doi: 10.1145/1866029.1866040

Liu, A., Guerra, S., Fung, I., Matute, G., Kamar, E., and Lasecki, W. (2020). "Towards hybrid human-AI workflows for unknown detection," in *Proceedings of The Web Conference 2020, WWW '20* (New York, NY: Association for Computing Machinery), 2432–2442. doi: 10.1145/3366423.3380306

Manam, V. C., and Quinn, A. J. (2018). "Wingit: efficient refinement of unclear task instructions," in *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Manam, V. K. C., Jampani, D., Zaim, M., Wu, M.-H., and J. Quinn, A. (2019). "Taskmate: a mechanism to improve the quality of instructions in crowdsourcing," in *Companion Proceedings of The 2019 World Wide Web Conference*, 1121–1130. doi: 10.1145/3308560.3317081

Marshall, C. C., and Shipman, F. M. (2013). "Experiences surveying the crowd: reflections on methods, participation, and reliability," in *Proceedings of the 5th Annual ACM Web Science Conference*, 234–243. doi: 10.1145/2464464.2464485

McDonnell, T., Lease, M., Elsayad, T., and Kutlu, M. (2016). "Why is that relevant? Collecting annotator rationales for relevance judgments," in *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 10. doi: 10.24963/ijcai.2017/692

Nguyen, A. T., Halpern, M., Wallace, B. C., and Lease, M. (2016). "Probabilistic modeling for crowdsourcing partially-subjective ratings," in *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 149–158.

Nouri, Z., Gadiraju, U., Engels, G., and Wachsmuth, H. (2021a). "What is unclear? Computational assessment of task clarity in crowdsourcing," in *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 165–175. doi: 10.1145/3465336.3475109

Nouri, Z., Prakash, N., Gadiraju, U., and Wachsmuth, H. (2021b). "iclarify-a tool to help requesters iteratively improve task descriptions in crowdsourcing," in *Proceedings of the 9th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Papoutsaki, A., Guo, H., Metaxa-Kakavouli, D., Gramazio, C., Rasley, J., Xie, W., et al. (2015). "Crowdsourcing from scratch: a pragmatic experiment in data collection by novice requesters," in *Third AAAI Conference on Human Computation and Crowdsourcing*.

Pickard, G., Pan, W., Rahwan, I., Cebrian, M., Crane, R., Madan, A., et al. (2011). Time-critical social mobilization. *Science* 334, 509–512. doi: 10.1126/science.1205869

Retelny, D., Robaszkiewicz, S., To, A., Lasecki, W. S., Patel, J., Rahmati, N., et al. (2014). "Expert crowdsourcing with flash teams," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 75–85. doi: 10.1145/2642918.2647409

Rosser, H., and Wiggins, A. (2019). "Crowds and camera traps: genres in online citizen science projects," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*. doi: 10.24251/HICSS.2019.637

Schaekermann, M., Goh, J., Larson, K., and Law, E. (2018). Resolvable vs. irresolvable disagreement: a study on worker deliberation in crowd work. *Proc. ACM Hum. Comput. Interact*. 2, 1–19. doi: 10.1145/3274423

Scholer, F., Kelly, D., Wu, W.-C., Lee, H. S., and Webber, W. (2013). "The effect of threshold priming and need for cognition on relevance calibration and assessment," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 623–632. doi: 10.1145/2484028.2484090

Sen, S., Giesel, M. E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., et al. (2015). "Turkers, scholars, arafat and peace: cultural communities and algorithmic gold standards," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work* & *Social Computing*, 826–838. doi: 10.1145/2675133.2675285

Sheshadri, A., and Lease, M. (2013). "SQUARE: a benchmark for research on computing crowd consensus," in *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, 156–164.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). "Cheap and fast–but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics), 254–263. doi: 10.3115/1613715.1613751

Sorokin, A., and Forsyth, D. (2008). "Utility data annotation with amazon mechanical Turk," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008, CVPRW'08*, 1–8. doi: 10.1109/CVPRW.2008.4562953

Tian, Y., and Zhu, J. (2012). "Learning from crowds in the presence of schools of thought," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 226–234. doi: 10.1145/2339530.2339571

Vakharia, D., and Lease, M. (2015). "Beyond mechanical Turk: an analysis of paid crowd work platforms," in *Proceedings of the iConference*.

Vandenhof, C. (2019). "A hybrid approach to identifying unknown unknowns of predictive models," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7*, 180–187.

Vidgen, B., and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: garbage in, garbage out. *PLoS ONE* 15:e0243300. doi: 10.1371/journal.pone.0243300

Vogels, W. (2007). *Help Find Jim Gray*. Available online at: https://www.allthingsdistributed.com/2007/02/help_find_jim_gray.html

Wang, F.-Y., Zeng, D., Hendler, J. A., Zhang, Q., Feng, Z., Gao, Y., et al. (2010). A study of the human flesh search engine: crowd-powered expansion of online knowledge. *Computer* 43, 45–53. doi: 10.1109/MC.2010.216

Wu, M.-H., and Quinn, A. J. (2017). "Confusing the crowd: task instruction quality on amazon mechanical Turk," in *Proceedings of the 5th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Zheng, Y., Li, G., Li, Y., Shan, C., and Cheng, R. (2017). Truth inference in crowdsourcing: is the problem solved? *Proc. VLDB Endow*. 10, 541–552. doi: 10.14778/3055540.3055547

Author Disclaimer: Any opinions, findings, and conclusions or recommendations expressed by the authors are entirely their own and do not represent those of the sponsors.

Conflict of Interest: ML is engaged as an Amazon Scholar, proposed methods are largely general across paid and volunteer labeling platforms.

The remaining 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/frai.2022.818491

How Personality and Communication Patterns Affect Online *ad-hoc* Teams Under Pressure

Federica Lucia Vinella[1]*, *Chinasa Odo*[2], *Ioanna Lykourentzou*[1] *and Judith Masthoff*[1]

[1] *Human Centred-Computing, Information and Computing Sciences, Utrecht University, Utrecht, Netherlands,* [2] *The School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, United Kingdom*

Critical, time-bounded, and high-stress tasks, like incident response, have often been solved by teams that are cohesive, adaptable, and prepared. Although a fair share of the literature has explored the effect of personality on various other types of teams and tasks, little is known about how it contributes to teamwork when teams of strangers have to cooperate *ad-hoc*, fast, and efficiently. This study explores the dynamics between 120 crowd participants paired into 60 virtual dyads and their collaboration outcome during the execution of a high-pressure, time-bound task. Results show that the personality trait of Openness to experience may impact team performance with teams with higher minimum levels of Openness more likely to defuse the bomb on time. An analysis of communication patterns suggests that winners made more use of action and response statements. The team role was linked to the individual's preference of certain communication patterns and related to their perception of the collaboration quality. Highly agreeable individuals seemed to cope better with losing, and individuals in teams heterogeneous in Conscientiousness seemed to feel better about collaboration quality. Our results also suggest there may be some impact of gender on performance. As this study was exploratory in nature, follow-on studies are needed to confirm these results. We discuss how these findings can help the development of AI systems to aid the formation and support of crowdsourced remote emergency teams.

Keywords: crowdsourcing, collaboration, social computing, personality, emergency response

Edited by:
Jie Yang,
Delft University of Technology,
Netherlands

Reviewed by:
Dalila Durães,
University of Minho, Portugal
Thomas Mandl,
University of Hildesheim, Germany

****Correspondence:***
Federica Lucia Vinella
f.l.vinella@uu.nl

Specialty section:
This article was submitted to
Machine Learning and Artificial
Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

Received: *19 November 2021*
Accepted: *05 May 2022*
Published: *27 May 2022*

Citation:
Vinella FL, Odo C, Lykourentzou I and Masthoff J (2022) How Personality and Communication Patterns Affect Online ad-hoc Teams Under Pressure. Front. Artif. Intell. 5:818491. doi: 10.3389/frai.2022.818491

1. INTRODUCTION

Situations that require working together, fast, and efficiently under pressure are on the rise, especially in an increasingly fragile global ecosystem (Schneider, 2011; Kretzschmar et al., 2022). From handling widespread geopolitical conflicts (Friede, 2022) to mitigating environmental disasters (Gay-Antaki, 2021), several organizations are investing in crowdsourcing intervention to aid large-scale mobilization of resources including emergency shelters and disaster-event detection (Pettet et al., 2022; Stephens and Robertson, 2022; Zhang, 2022). Likewise, virtual teamwork enacted in high-urgency, high-stress tasks is on demand. Grassroots social engagement [i.e., Covid-19 pandemic hackathons (Colovic et al., 2022)], incident response squads (Palen et al., 2007), community response teams, and on-call software solution teams (Anderson, 2020) are all examples of ongoing large-scale collaborative efforts. Emergency teams are devolving into technology, and the internet, in particular, to enforce the timely resolution of complex problems within limited time frames, often under stress, and potentially with collaborators who have never worked together in the past. The benefits of working virtually and remotely are evident as shown by the thriving field of telemedicine with remote surgical teams aiding medical centers in coping with widespread pandemics (Etheridge et al., 2022). Nevertheless, little is known about the factors that can make or break such teams. In this study, we attempt to answer questions such as: *What are the personality characteristics that render high-stake online teams successful? Which skills, abilities, or socio-cultural elements are essential to consider while forming these teams? Are there any particular communication patterns that can serve as early signals of effective teamwork under stress?* Answering these questions is crucial to leverage available resources and intellect in critical situations. Although group research has since long investigated the effect of factors including personality, knowledge, skills, or socio-cultural facets on virtual teamwork (Kichuk and Wiesner, 1997; Krumm et al., 2016), few studies examine these characteristics on the specific problem of online collaboration strained by external—psychological or time-related—aspects.

Teams performing in rapid response environments do not perform similarly to "normal" teamwork settings. They are under pressure from the high-demand context under which they operate. The time-bounded nature of the task increases the chances of failure (Driskell et al., 2018). Characteristics of team performance in rapid-response, high-stress contexts are team members' ability to work in a team and personality traits (McManus et al., 2004; Subramaniam et al., 2010). However, to date, studies on high-stake teams focus either on emergency professional teams, crowd participation in emergency response, or the collaboration between these two groups without considering the aspect of team formation at the crowd level. Our study observes **remote, ubiquitous, online, and *ad-hoc* crowd teams** instead of traditional emergency response offline teams with specialized individuals (Chen et al., 2008). We deem the crowd, alongside teamwork emergency response, as the two most relevant aspects of this research, as we analyze and report properties contributing to successful outcomes under situations of stress and ambiguity. Furthermore, we examine the relationship between personality, socio-cultural elements, and communication patterns on the one hand, with team performance and satisfaction on the other, in the context of *ad-hoc online teams in rapid-response, high-pressure tasks*.

1.1. The Task: A Virtual Maze for Remote Crowdsourcing Emergency Teamwork

To study participant interactions in *ad-hoc* teams of strangers under pressure, we turn to crowdsourcing, and a custom-made task. Our task is inspired by the "Keep Talking Nobody Explodes" (Knuth, 2021) puzzle video game. Participants work in dyads, and their common mission is to defuse a bomb that is placed within a maze, by combining information that is unique to each one of them. One participant is assigned the role of the "Defuser": they can "walk" inside the maze toward the bomb and defuse it, but they do not know where the maze walls are. The other participant is assigned the role of the "Lead Expert": they have the map of the maze but they cannot walk in it. The Defuser and the Lead Expert must exchange information and actions, to defuse the bomb within a limited amount of time. The task has been designed to have the same critical characteristics as actual emergency response tasks, namely a high-demanding environment, enforced role division, performance pressure and stress.

1.1.1. High-Demanding Environment

Instances of crisis constitute a large part of what emergency teams have to deal with and radically define their functional and structural properties. Demanding environments have critical requirements with tangible consequences for poor performance (e.g., accidents, errors, stress). By portraying the element of urgency in the form of a virtual bomb and increased time pressure (Bell et al., 2018) we focus on a single objective—reaching the bomb on time—and deliver the results of a study task that is critically cooperative and built for productive communication. In our setting, virtual crowd teams must deliver innovative solutions and deliver them quickly. The typical environmental constraints of high-demanding tasks (time, urgency, risks) command for independent, stable, role-defined teams sharing mutual trust, values, and focus. As we reduce and inter-mediate communication through digital means, we impose an even further reliance on mutual objectives, well-defined roles and obligations, effective communication, and commitment.

1.1.2. Enforced Role Division

During cases of emergency, each team member has a distinct and specific role to play (Baldwin and Woods, 1994), which is typically a-priori and externally defined. Emergency and periods of crisis often create the need for established protocols of interaction respective to each part (Harrison and Connors, 1984). Although role division is typically fixed for these response units (e.g., medical, logistic, security, public relations, etc.), it must nonetheless be adaptable when facing unpredictable outcomes. By assigning strangers to pre-defined roles, we replicate a scenario where team roles are agreed upon yet flexible and interposed. Through well-defined roles and responsibilities, we evaluate the matching capabilities of crowd workers and investigate what are the constituents that fundamentally determine the execution of role-based virtual teamwork emergency response.

1.1.3. Performance Pressure and Stress

Prior work has shown that users involved in games such as the crowdsourcing task exhibit various forms of stress (Sabo and Rajčáni, 2017) and heightened emotional states (Hart et al., 2018). These teams are more susceptible to allostatic load, i.e., the process of "wear and tear" experienced by team players facing stressful conditions (Davaslioglu et al., 2019). Regarding the definition of stress, there are two kinds of stressful conditions and stressors (Ma et al., 2021). One definition follows the general assumption that a stressor (the triggering factor) negatively affects the person by degrading performance; the other sees stress as a challenge that improves performance and individual

TABLE 1 | Positive and negative facets of the BIG-5 personality traits (Neuman et al., 1999).

Big five traits	Positive facets	Negative facets
Extraversion	Social, talkative, assertive, active	Retiring, sober, reserved, cautious
Agreeableness	Good-natured, gentle, Cooperative, hopeful	irritable, suspicious, uncooperative, inflexible
Conscientiousness	Self-disciplined, responsible, Organized, scrupulous	lacking self-discipline, irresponsible, Disorganized, unscrupulous
Emotional stability	Calm, enthusiastic, Poised, secure	Anxious, depressed, Emotional, insecure
Openness to experience	Imaginative, sensitive Intellectual, curious	down-to-earth, insensitive, simple, narrow

gains (Zhang and Lu, 2009). In this research, we stripped the task from several elements of the original video game with the intent to transverse from multiple sources of hindering stressors [that increase environmental demands and exceed the available resources (Salas et al., 1996; Gardner, 2012)] to a unique challenge to inspire and motivate collaborators. Finally, virtual teams experience stress differently than offline ones as they tend to experience lessened social support (Su et al., 2012) which exacerbates predispositions to stress and anxiety (Tarafdar and Stich, 2021). For this reason, even though we adjusted the task to limit encumbrance, we still regard the individual and team response to a stressful task as the determining factor for whether personal characteristics and/or team compositions help handle the challenge successfully.

By engaging the players in this high-pressure challenge, we examine whether personality characteristics (Conscientiousness, Extraversion, Neuroticism, Agreeableness, and Openness) may make individuals more prone to cooperation under time pressure. We further evaluate which, if any, combination of personalities results in better than average team performance. Similarly, we examine whether additional factors such as the participants' socio-cultural background affect their actual ability to work together and their satisfaction with teamwork. Understanding the crowds perception of the collaboration (and not only performance) will help the development of AI agents to support their needs—and not only effectiveness—in times of crisis. Additionally, perceptions on the collaboration may provide insights into why certain teams are more effective than others, and what teams may be willing to work together again on the next task. Thanks to the heterogeneous data gathered during the experiment, we look at the dyadic communication to unravel indicators of a given team's potential to cope with a high-demanding task under time pressure.

A focus of this research is the impact of participants' personality on *ad-hoc* online teamwork, that is crowd-sourced, brief, and under pressure. We use the Big Five personality model (Goldberg, 1990), also known as the Five-Factor model, to model and comprehend the relationship between crowd workers' personality traits and their disposition for online teamwork in emergency contingencies. We selected the Big Five model as it is most commonly used for personality analysis [e.g., Highhouse et al., 2022; Ikizer et al., 2022; Mammadov, 2022] and for artificial intelligence systems that automatically adapt to personality [see (Smith et al., 2019) for a review of personality models used for personalization in persuasive technology, intelligent tutoring systems and recommender systems]. Additionally, many validated instruments exist to measure the Big Five traits, including the brief version of the Big Five Personality Inventory (Rammstedt and John, 2007) which we use in this paper. The Big Five model distinguishes between 5 traits[1], each of which has multiple facets (see **Table 1**)

1.2. Research Scope: Human Factors for AI Intervention in Crowdsourcing Emergency Response Teams

As work shifts to increasingly digitized spaces and connections between collaborators are made broader by mobile and ubiquitous computing, we consider evaluating ways to channel these resources to help remote, crowdsourced emergency teams. Identifying attributes and interactions used in emergency crises can help organizations and research improve upon methods for remote communication. Our knowledge of characteristics that contribute to virtual emergency response teamwork can inform artificial intelligent systems in assessing whether and how an individual can be part of a response unit with limited time and resources, and also, if multiple possible workers and tasks exist, who to use for the emergency response teams.

The rest of the paper is organized as follows. Section 2 presents and discusses related work, including an overview of traditional teams under pressure and crowdsourcing efforts in this domain, as well as the study hypotheses. Section 3 describes the study design, including participant sample and task design. Section 4 describes the metrics used to capture participants' demographic characteristics, Big Five personality traits, and ability (prior experience and self-perceived ability), as well as the metrics of teamwork, namely: collaboration quality and communication patterns. Section 5 presents the results. In Section 6 we discuss the implications of this work, its limitations, and possible extensions for the future. Finally, section 7 concludes the paper with key findings and closing remarks.

2. RELATED WORK

2.1. Teams in Classical High-Demand, Time-Pressing Settings

2.1.1. Operational Setting and Problem Scope

Significant research effort has been placed over the years on teams that need to perform in situations that require spontaneous, *ad-hoc* decisions and short-term planning, to resolve ambiguous or uncertain events, and where the consequences of failure are significant (Reuter et al., 2014). The scope of the problems that such teams are called to deal with is broad. It can include responding to natural disasters, like floods, hurricanes, and fires, but also managing crises (King, 2002), such as

[1] Emotional Stability is often replaced in literature by its opposite Neuroticism.

terrorism events (Longstaff and Yang, 2008), events occurring in long-duration spaceflights (Salas et al., 2015), nuclear plant control rooms (Stachowski et al., 2009), or situations taking place in a military context (Driskell et al., 2014). It can also include more benign everyday workplace settings, such as on-call software teams dealing with organizational incidents, like security or service failure events (for example the recent Google outage (Bergen, 2020), journalist teams for the immediate coverage of unexpected events (Archibold, 2003), but also short-term project teams (Galbraith and Lawler, 1993) and task forces (Hackman, 1990). Their size can vary, from dyads and triads (Foushee, 1984), to dozens (Helmreich, 1967), to twenty or more (Stuster, 2011).

2.1.2. Differences From Normal Teams

What separates these teams from teams in "normal" settings, is the extreme, atypical environment within which they operate, which overall entrails time pressure, high levels of risk, increased consequences for poor performance (Driskell et al., 2018), no previous work experience with one another, and the need to perform their task almost immediately on team formation (Mckinney et al., 2005; Mendonça, 2007). Harrison and Connors (1984) use the term exotic environment to describe a work setting that is marked by hostile environmental demands, restricted working conditions, isolation from those outside the setting, and confinement and enforced interactions for those inside it. Using the related term extreme environment, Bell et al. (2018) add that these settings are also characterized by limited time to finish the task. Performance pressure and severe consequences for ineffective performance are also characteristic of these settings, and this pressure can act as a double-edged sword that can lead the team to outstanding performance, or cripple it Gardner (2012). The tasks that teams in these settings must solve are usually characterized by ambiguity and urgency (Yu et al., 2008; Stachowski et al., 2009).

2.1.3. Factors Affecting the Success of Emergency Teams

Which factors determine team success in this high-demand, high-stress environment? *Skill* and expertise are the primary factors. Teams traditionally trained as emergency response units rely on the specialized expertise of the stages of the incident response and carry insider knowledge of the organizational policies, their obligations, the communication channels, and the tools supplied by the hiring organization. Thereof, the effectiveness of traditionally formed emergency response teams relies to a great extent on the level of preparedness and competence of the hiring body (or authority) that trained and assembled them, with multiple historical incidents providing evidence for the need for precise training programs and hiring criteria (Alexander, 2003). Examining command and control teams, Ellis et al. (2005) find that team members with higher training demonstrated greater proficiency in planning and task coordination activities, as well as in collaborative problem-solving, and communication. The study also found that it is the knowledge competencies of the team member with the most critical position that benefited the team the most.

The second factor of interest is the allocation of *roles and authority*. A prominent characteristic of typical high-stake teams, such as STAts (swift-starting action teams), is that they comprise experts (Mckinney et al., 2005) with specific roles and responsibilities. Multiple studies confirm the value of stable role structure in the division of labor and in enhancing the predictability of team interactions, allowing each team member to know what to expect from their teammates in critical situations (Hackman and Morris, 1975; Stachowski et al., 2009). The reason is that misunderstandings or disagreements about authority and role accountability (especially non-desirable roles like clean-up) may lead to team conflict, especially in the presence of unprecedented emergency response tasks (Quarantelli, 1988; Weick, 1993). The meta-analysis of De Wit et al. (2012) further confirms the negative relationships between process and role conflict, and team results such as cohesion, commitment, and performance. On the other hand, flexibility, the ability to improvise, and entrusting functional requirements to determine roles, rather than relying on titles may also be of benefit (Briggs, 2005; Mendonça, 2007). A highly defined role structure with clear roles seems to benefit more tasks that are structured. On the contrary, a flatter structure may be better for ambiguous tasks for which no apparent solution can be easily found (Worchel and Shackelford, 1991) (such as the task of responding to the 2001 World Trade Center attack Mendonça, 2007).

Personality is another prominent factor affecting the success of high-stakes teams, in line with the broader personnel selection literature which indicates that if relevant personality factors are identified for a specific job, future performance can be predicted (Borman et al., 1980). Using the occupational personality questionnaire to study the emergency command ability of offshore installation managers, Flin and Slaven (1996) finds significant correlations between command abilities in critical situations and certain personality elements. From their results, it appears that the highest-rated performance came from those who (a) like to take charge and supervise others (high score on controlling), (b) consider themselves to be fun-loving, sociable, and humorous (high score on outgoing), (c) are less interested in analyzing human behavior (low score on behavioral), (d) are more interested in practical than abstract problem solving (low score on conceptual), and (e) prefer to make decisions quickly rather than take time to weigh up all the evidence (high score on decisive).

Flin and Slaven (1996) contribution, however modest in size, is only pertinent to emergency command responsibilities and applicable only within a specific type of organization (offshore installation managers). Other researchers have focused on the possible existence of a "rescue personality," in multiple additional domains where emergency services and occupational stress are pivotal. Kennedy et al.'s (2014) research on how personality influences the workforce decisions of emergency nurses reveals that certain traits matter more than others. High Extraversion, Openness to experience, and Agreeableness were especially common amongst emergency nurses. Extraversion was also present among emergency department senior medical staff (Boyd and Brown, 2005) as part of the controversial

ENTJ (Extrovert, Intuitive, Thinking, Judging) personality type[2] (Myers, 1962).

Partially supporting these findings is the work of Wagner et al. (2009) on the personality traits of paid professional firefighters. Although high Conscientiousness was not a determinant factor in this vocational role, Extraversion had significance. Certain personality traits seem to cluster under particular types of emergency professions; the differentiation between correlation and causality between these two variables is not always easy to untangle. Feelings of anxiety and insecurity, as well as heightened levels of Neuroticism and Openness, were seen to be most likely the results, and not the cause, of the repetitive exposure to experiences of loss and distress (Pajonk et al., 2011). By broadening the sample to the general public (virtual crowd), we aim at decoupling the effects that a specialized profession could have on one's propensity to emergency response.

Finally, certain *interaction patterns* are useful predictors of whether an *ad-hoc* team that has been brought together for immediate task performance will succeed or not, in classical emergency response teams. Although swift-start teams have little time to build their group processes before starting to work on the task, it is also known that team routines get established early in the team's lifecycle. The same initial interactions have an effect on subsequent communication and norms (Gersick and Hackman, 1990). The study of Zijlstra et al. (2012) reveals that there are certain early patterns of communication that distinguish effective from less effective teams. Specifically, they find that effective teams engage in communication that is more stable in duration and complexity, more balanced, and less monopolized by a single participant compared to inefficient teams that exhibit frequent mono-actor patterns, consisting of a single team member posing and answering their questions and commenting on their observations. They also found that efficient teams exhibit more reciprocity and trust, with the team members engaged and in the same direction of action toward the task goal. The presence of trust as a crucial factor is also highlighted (Wildman et al., 2012). The study of Waller et al. (2004) reveals that efficient teams in non-routine situations focused their actions on information collection and task prioritization. Finally, Kanki et al. (1991, 1989) complement the above by showing that the communication of effective swift-start two-person crews focuses on immediate task execution, expressed as low-complexity, straightforward action statements, and is less focused on other non-standard communication.

Although classical rapid-action teams are widely studied, these literature findings do not necessarily translate to online crowd rapid-action teams. Traditional emergency teams comprise highly trained professionals with a shared understanding of the crisis domain, and often a shared loyalty to an organization. In contrast, crowd teams mainly consist of non-experts, and they are more volatile and heterogeneous regarding the motivators that draw their members to the particular task. Considering the multiplication and globalization of the events that require swift action, it is likely that in the future, we will need to turn more and more to crowd workers and volunteers to form *ad-hoc* online teams that can deal with high-stake situations under pressure. In this light, the extensive study of classical rapid-action teams can provide us with the first grounded indications of specific parameters to look at to identify predictors of successful team formation in online crowd action teams. Given that in a crowd setting, the allocation of roles is likely to take place based on arrival and availability, in this work, we focus on the parameters of personality and communication patterns as predictors of forming a successful crowd team to tackle unforeseen situations under time pressure.

2.1.4. Onsite and Offsite Emergency Response Teams

The history of emergency response teams—and more broadly of emergency preparedness—is essentially as old as societal and humanitarian threats. For as long as emergencies have affected human lives, societies have found collective ways to organize efforts to mitigate, prepare, respond, and recover from the aftermaths of crises. Emergency preparedness programs have evolved along with societal changes and technological advancements. Notable historical events such as the first world war brought national societies to unify and strengthen their approaches to natural, intentional, and accidental disasters (Herstein et al., 2021). The International Federation of Red Cross and Red Crescent Societies is one of the most prominent products of global pursuits unifying volunteer networks, community-based expertise, and independent advisers into standardized practices (London, 1998). As emergency response evolves, emergency response teams reshape ways to communicate and function in an era of accelerated technological progress.

Formerly, emergency teams operated face-to-face and on-site in response to environmental disasters (Brennan and Flint, 2007), war conflicts (Abdul-Razik et al., 2021), and epidemics (Leach et al., 2022). With the broadening digitization of services, society is increasingly reliant on technology for its functioning. The so-called information era entails the vast market of the internet of things, software, and the worldwide web to enable widespread financial and data transactions (Stehr, 2001). Technological dependency is making us faster and smarter and, at the same time, more vulnerable to novel threats (e.g., malware attacks, identity theft, financial fraud, security breaches, etc.). Emergency response teams not only must face novel and extensive digital threats but must also learn to leverage the resourcefulness of recent technology [ubiquitous computing (Smirnov et al., 2011), robotics (Kawatsuma et al., 2012), simulations (Kincaid et al., 2003), smart sensors (Abu-Elkheir et al., 2016), and social media networks Potts, 2013] to strengthen their outreach and preparedness.

Overall, the vast majority of emergency response teams operate in a hybrid fashion combining onsite support with online offsite communication. Some others divide efforts between online and face-to-face tasks depending on the phase of the response (i.e., mitigation, preparedness, response, and recovery Brennan and Flint, 2007). Relevant to our research is the

[2] studies have been conducted on construct MBTI validity and test-retest reliability (including a meta-study by Capraro and Capraro (2002) which showed good results), others have argued that there are scientific limitations to these studies, the use of MBTI, and its underlying theory (e.g., Boyle, 1995; Pittenger, 2005; Stein and Swan, 2019).

pertinence of virtual communication channels in the large-scale crowdsourced emergency response domain that is typically remote, collaborative, and online. To define our target group, we firstly identify general characteristics that, in the classical sense, differentiate between onsite and offsite emergency response teams. Although the two domains share very similar objectives and attributes such as organizational culture, expertise, team structure, communication, and teamwork (Leach and Mayo, 2013), since their capabilities and duties differ, some of these attributes are more imperative than others. In the following subsections, we introduce two representative attributes critical for each teamwork domain.

2.1.4.1. Onsite Emergency Response Teams

Two prominent attributes of onsite teams are **experience** and **coordination**. Teams working onsite are usually part of rescue operations (Chen and Miller-Hooks, 2012) and disaster relief (Bjerge et al., 2016) that require the participation and coordination of experts. These include fire and rescue services and police forces, commercial entities, volunteer organizations such as the Red Cross, media organizations, and the public (Yang et al., 2009). The need for distinct expertise requires teams to develop and apply specialized knowledge. Onsite emergency response experts can hold intelligence on chemical properties, procedures for reporting emergencies, fire and protective equipment, decontamination, and evacuation gained through training, experience, and/or formal education.

Without qualified knowledge and standardized procedures, onsite emergency response teams would fall short of promptly and accurately addressing ongoing crises. Equally important is coordination among experts as onsite emergency must successfully distribute superintendence and responsibilities between diverse professionals for effective prevention, preparedness, and response to emergencies. In their work on coordination in emergency response management, Chen et al. (2008) developed a life-cycle approach with three distinct sets of activities on the timeline continuum (pre-incident phase, during incident phase, and recovery phase). The cycle closes after de-briefing and when actionable items are learned from the intervention and incorporated into the plan to affect future preparedness (Chen et al., 2008). The same authors identified several elements of coordination such as activities, coordination objects, and constraints that differ between phases and between cultural, political, regulatory, and infrastructural properties of emergency response.

2.1.4.2. Offsite Emergency Response Teams

Two distinguishing attributes of offsite remote emergency response teams are **communication** and **sensemaking**. While onsite teams converge in rescue operations and disaster relief, remote offsite emergency response teams outreach and distribute resources. Known crises overseen by offsite emergency response teams are air-traffic control (Hughes et al., 1992), subway crisis management (Heath and Luff, 1992), and emergency response call centers (Normark, 2002; Pettersson et al., 2004). Although clear roles are important in these teams, clear communication is of the essence. Depending on the kind of interaction (e.g., serendipitous, inbound, and outbound Landgren and Nulden, 2007), and the referent (e.g., non-experts' communication, situation update, situational awareness, services access assistance Velev and Zlateva, 2012), clear communication and interaction protocols fundamentally determine the interaction mediated by computer systems for offsite rescue teams.

Through clear communication, offsite emergency response teams can harvest sensemaking. This is the collection of actions that make the situation understandable and that prevent an escalation of the emergency (Landgren and Nulden, 2007). Sensemaking has properties such as identity construction, retrospection, enactment, social reactions, dynamism, environmental cues, and plausibility (Muhren et al., 2010). The importance of sensemaking in a remote emergency context is ever so apparent due to the practical constraints that teams experience as they communicate remotely. According to Weick (1993), most shortcomings from failed emergency responses are due to a deficiency in sensemaking (or contextual rationality). Weick (1993)'s work uncovers four potential sources of resilience that make *ad-hoc* groups less vulnerable to disruption of sensemaking. These sources are (i) improvisation, (ii) virtual role systems, (iii) the attitude of wisdom, and (iv) norms of respectful interaction. Weick (1993) analyses the dynamics of role structure and sense-making occurring in the historical Mann Gluch disaster. The incident served as an example of what needs to be re-examined about temporary systems, structuration, non-disclosed intimacy, inter-group dynamics, and team building (Weick, 1993), especially important for offsite emergency response operations.

The design of computer-mediated emergency response also needs to be informed by an understanding of the cognitive processes involved in responding to unanticipated contingencies (Mendonça, 2007). These cognitive factors, defined by Mendonça (2007), are directly linked to the specificity of emergence management and its characteristics of rarity, time pressure, uncertainty, high and broad consequences, complexity, and multiple decision making. Besides, computer-mediated emergency response teams are much more predisposed to incorporate the output of citizen convergence (Schmidt et al., 2018) into their work than traditional onsite rescue teams. However, as developments in online informational convergence change the remote domain of rescue operations, citizens and crowds are bringing novel paradigms. These include unfamiliar team members, ill-defined tasks, fleeting membership, multiple and conflicting goals, and geographically distributed collaboration (Majchrzak and More, 2011). In the following section, we explore the topic of crowdsourcing for emergency response.

2.2. Crowdsourcing for Emergency Response

2.2.1. Emergency Response Through Individual Crowd Contributions

Crowds are increasingly involved in response to emergencies. The characteristic of emergency response crowdsourcing is the short-lived engagement in the task. Crowds' contributions consist of primarily individual, one-time, and remote interactions. This "long-tail" of contributions is a well-observed phenomenon in most content-oriented online communities (Shirky, 2008). The role of these one-time crowd users is important when it acts as a fast and ubiquitous response to urgent, environmental and social crises (hurricanes, terrorist attacks, widespread fires, large oil spills, etc.) (Heinzelman and Waters, 2010; Yuan and Liu, 2018; Chau, 2020), protest movements (Elsafoury, 2020), but also activism (Farkas and Neumayer, 2017; Lee, 2020) and civic participation (Hemphill and Roback, 2014; Mitchell and Lim, 2018). In critical scenarios of this kind, the crowd is intended as a manifold social tool by servicing as a reporter, social computer, sensor, and executor of both micro and macro-tasks.

Several theoretical studies propose system models and features designed to facilitate the positioning of the crowd as the leading resource for emergency management. In the domain of communication technologies for health care Hossain et al. (2017) suggest benefiting from the users' social contacts to trigger a faster response, or to make the most of crowdsourcing attributes—such as collaboration and tournaments—to attract the right crowd for the job. From a complex systems perspective, Song et al. (2020) propose harnessing the self-organizing operation mechanisms of crowdsourcing for efficient disaster governance. In the context of natural disaster management, Ernst et al. (2017) propose hybrid systems that rely on the remote coordination of volunteers to collect location-dependent information, which in turn can support emergency managers making quick but solid decisions. Elsafoury (2020) propose another hybrid feature, this time combining machine learning with crowdsourcing to rapidly detect protest repression incidents through social media.

Specific crowdsourcing tools and platforms address emergencies. Poblet et al.'s (2013) review indicates that these platforms belong to two main categories, namely: (i) data-oriented, and (ii) communication-oriented. The first category concerns tools developed for the intensive aggregation, mining, and processing of data gathered through the crowd. The second category aims at supporting communication between crowd users and disaster management systems by allowing seamless interaction between them. The platform "Ushahidi" (Okolloh, 2009) is one example of a crowd application designed to decentralize the support of volunteers for the report of violence in Kenya, by collecting sensitive reports, organizing rapid response actions across multiple agencies, documenting ongoing changes, generating automatic alerts from under updates and visualizing data streams in real-time.

In another example, several digital volunteer organizations (Standby Task Force, Humanity Road, and Open Crisis) have integrated social media monitoring in their systems when cooperating with other humanitarian bodies in disaster relief operations (Poblet et al., 2013) Poblet et al.'s (2013) review of crowdsourcing tools for disaster management offers an extensive list of crowdsourcing tools, including online platforms and mobile applications across the globe. Aside from those tools that support response and recovery-based only efforts, others, such as ArcGIS (Allen, 2011), Sahana (Careem et al., 2006), OpenIR (Ducao, 2013), and CrisisTracker (Rogstadius et al., 2013), provide support for mitigation and crisis preparedness. These tools pivot around the crowd for achieving great humanistic and environmental causes while leveraging the strength of geographically dispersed collaboration.

However, despite the growth of several initiatives and digital platforms designated to facilitate crowd intervention in emergency response, these initiatives are primarily based on individual contributions, without taking advantage of team dynamics that can arise among the crowd participants. This lack of communication, either due to team conflict (Yeo et al., 2018), or unfitness of the tools (Dilmaghani and Rao, 2006), makes crowdsourcing efforts less efficient, which often fail to address the event at hand, either as standalone initiatives or as supporting capacity to expert emergency management (Heath and Palenchar, 2000). Beyond the subject of crowdsourcing for emergency response, other team categories are also relevant to our research on *ad-hoc* crowd team formation. Action teams, rapid response teams, and citizen science, to name a few, are groups formed through the crowd and behave similarly to *ad-hoc* teams. Similar entities could benefit from system improvements addressing better team formation and communication strategies adopted from a better understanding of team dynamics in stressful situations. In the following subsection, we elaborate on existing—albeit early—efforts that seek to involve the crowd in formations and groups.

2.2.2. Crowd Cooperation for Emergency Response

Aside from individual crowd contributions, a few studies have looked into facilitating communication among crowd members to respond to and manage unexpected events. Providing people with communication channels can help them gain a broader view of the event they need to deal with (Perez and Zeadally, 2019), and better coordinate their efforts (Martella et al., 2017). Song et al. (2020) analyzed a total of twelve international case studies of crowdsourcing and natural disaster governance. They denote that, across all of these instances, the crowd manifested (at least at some level in their response mechanisms) self-organizing properties that lead its individuals to form collaborative ties spontaneously. It suggests that the multi-directional relationship between the crowdsourcing platforms, the initiators, and the contractors, while not strictly guided, triggers the formation of functional teams that act as active response units. Under this instance, the crowd forms *ad-hoc* groups as the emerging outcome of community disaster resilience (Song et al., 2020). As long as collaboration is advantageous in emergency response and time management remains vital in real-life crises, boosting the efficacy of crowd participation starting from the level of team formation can get teams closer to their desired outcomes.

Many combinations of individual traits add up as building blocks for the entire social entity that is the team. Assuming that the single characteristic is, at least in principle, an optimal fit for the task, the way it interacts with the rest of the teammates' features is equally relevant. Personality clashes are present in virtual team interactions just as in traditional face-to-face cases. Following Van de Ven et al. (1976) definition of teams as "groups becoming more effective over time," Salehi et al.'s (2017) work on stable crowd teams recognizes familiarity as the utmost important factor that enhances team performance. However, familiarity is a variable that cannot always be factored in when teaming up with individuals part of a virtual crowd, who are often sporadic contributors. Therefore, while familiarity in crowd teams has its tangible benefits (Salehi et al., 2017) for more stable tasks (like creative ones), relying on team familiarity to form effective crowd teams is not always feasible for short-lived, unpredictable, and mutable tasks.

For relatively short-lived assignments, the distribution of personality types matters more for the success and the establishment of trust in crowd teams than the pervasiveness of one specific type. Lykourentzou et al.'s (2016) work on crowd teams shows that balancing personality traits not only leads to significantly better performance on collaborative tasks but also reduces conflict and heightens the levels of satisfaction and acceptance. Holistically, when considering the impact of personality distribution in crowd teams, aspects other than personality traits play an often overlooked yet fundamental role. As Lykourentzou et al.'s (2016) noted: *test Personality could also be examined with regards to task type. For example, competitive tasks (like ideation contests among competing crowd teams) may amplify clashes within imbalanced teams, more than collaborative tasks*." We aim to uncover the relevance of personality, communication, and other factors in a virtual emergency response task. Unlike other studies (Floch et al., 2012; Vivacqua and Borges, 2012; Ernst et al., 2017) evaluating crowd emergency response as a collective and self-organized effort, we propose a team-specific approach to the formation of crowd emergency units that strongly connects with theories and models of teams composition, and assembly and team science (National Research Council, 2015).

Closing, most crowdsourced initiatives for high-stake, high-pressure tasks rely on individual contributions. Few works use some form of teamwork to coordinate crowd participants' efforts spontaneously and not according to a systematic approach or criteria. The formation of crowd emergency teams according to a set of characteristics with known expected effects could help these teams experience less interpersonal conflicts, establish team cohesion faster, and increase the teams' chances of success. In this work, we systematize online team formation for high-pressure tasks. We closely investigate the effects of personality and communication patterns, contributing to such teams' success and helping harness the crowd's potential better in emergency response.

3. STUDY DESIGN

Many factors may impact whether teams collaborate well and achieve their goals in an emergency response task. These include the demographics and personality of team members (both at the level of individuals and aggregated over the team), and the communication patterns used. This study explored which factors matter for team success and perceptions of collaboration quality. Given the many factors and output measures considered, the study was exploratory in nature, with the aim to gain initial insights into what matters and in which way, to be tested further in follow on studies.

3.1. Sample

120 Amazon Mechanical Turk workers (41 female, 78 male, 1 prefer not to say) participated. The task duration was approximately 20 min. Most participants were of U.S. (67 users) and Indian nationalities (51 users), one participant was Irish and another one was British. The majority had College (87) or Postgraduate degrees (15), while some had either some college education (9) or High School (9). Most were between 30 and 49 years of age. For an overview of the demographic data of the sample see **Table 7**.

3.2. Compensation

The participants received a base reward of $3, and a bonus reward of $3 if the challenge was completed successfully. The base pay was based on current fair crowd work compensation practices, whereas the bonus pay matched the base pay to double the reward for those teams that defused the bomb on time. The payment was weighted against the hourly rate or AMT workers as reported in Hara et al. (2018). In selecting the payment amount, we took into account three considerations from the literature (Olson and Kellogg, 2014; Lykourentzou et al., 2016). First, the payment had to conform to the community standards of the crowdsourcing platform so as not to bias the quality through workers who would accept low wages or workers who would only choose the task purely for its high compensation. Second, this payment had to cover the task duration. Thirdly, it took into account the demographics of the target worker population (minimum wage).

We recruited through the Amazon Mechanical Turk (AMT) Human Intelligent Task (HIT) platform. AMT was chosen for its breadth of crowd workers and its abundant labor availability, which is estimated to be no less than 2K workers at any given time, and over 100K workers overall (Difallah et al., 2018)[3]. No pre-selection was required to participate in the task. We intended to attract a large variety of participants, regardless of differences in background. The absence of pre-selection criteria may have influenced participants' written English, a limitation discussed in Section 6.2.3. Finally, the HIT itself contained information about the reward, the duration of the task, and a short description of the cooperative game.

3.3. Task Design and Setting

Although the task was artificial it was designed as an analog setting enacting the key characteristics of the high-demand, high-pressure environments that we are interested in. These include:

1. **Simulated element of physical danger**. The consequence of the team failing to navigate the maze is a bomb exploding. Although participants were aware that they are playing a game, the element of physical danger, even an enacted one, alters their perception, with possible effects on the way they process information, coordinate their efforts, and discuss (Kamphuis et al., 2011).
2. **Pre-determined team roles**. The presence of these roles enables stable and predictable group interactions (McMichael et al., 1999) instead of relying upon the slower and

[3] AMT worker's population is composed primarily of Indian and American nationalities, followed by Chinese, British, and Philippino (Difallah et al., 2018). The gender is slightly predominantly female within the American sample and more male in other countries (Difallah et al., 2018). Its population average age is less than the world population average, as most AMT workers were born after the 1990's (Difallah et al., 2018).

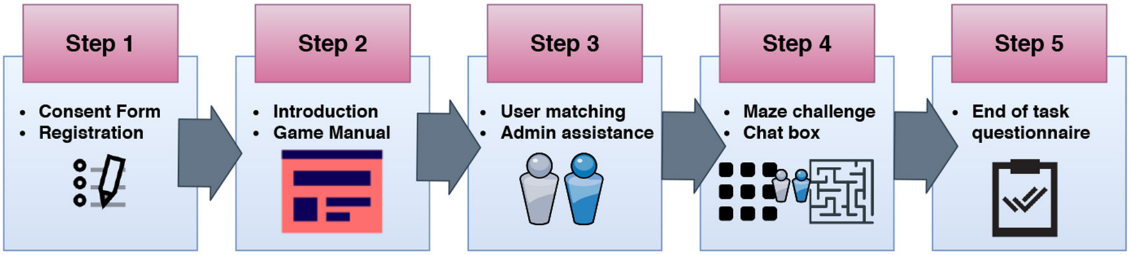


FIGURE 1 | System overview with the five steps of the study design. After registration, users arrive at an introductory page with relevant information about the task, and then they are matched in dyads on a first-in-first-out basis. Each team then proceeds to their dedicated virtual room where they cooperate to defuse the bomb in the maze within a given time frame. Finally, they fill out a questionnaire about their abilities and perceived collaboration quality.

autonomous differentiation of team roles (Belbin, 2012), which cannot always happen in circumstances of emergency. Predefined role-playing exercised control over one's limited access to information, which symbolizes the relationship between an overseeing entity (in our case, the Lead Expert) and an operative agent (in our case, the Defuser). Furthermore, similar to real-life action teams, team membership symbolizes work shifts (Zijlstra et al., 2012). It represents the random assignment of roles on a first-come-first-served basis. Similar to emergency response teams, this approach creates teams with little time to explore personal similarities and differences or to go through classical team development processes (Tuckman and Jensen, 1977; Lacoursiere, 1980).

3. **Stress and increased consequences of failure**. The novelty of the task, alongside its short duration, positions the crowd participants in a situation similar to emergency management scenarios. Here, the users need to act decisively within tight time schedules, often only with access to incomplete or difficult to decode information (Carver and Turoff, 2007). It means that the participants (a) absorb information rapidly, (b) judge by doing, (c) decide on the spot, (d) deal with the event with little preparation. Users are aware that their actions, if wrong, will cost them (and their teammate) reasonably significant retribution (in this case monetary) (Driskell et al., 2018). The combination of elements, namely: high-stake, time-constrained, fractional information, and role inter-dependency, makes this particular task a reasonably stressful one. More so, the original game "Keep Talking Nobody Explodes" has been utilized as a tool by past research to assess the effects of realistic stress on behavioral and physiological responses of participants (Sabo and Rajčáni, 2017; Lee and Jung, 2020). These studies confirm that controlled environments of this sort can correctly reproduce similar stress levels of more realistic scenarios, thus inducing stimulus-response events—such as temporary homeostatic changes and speech variations— that signal increased stress.

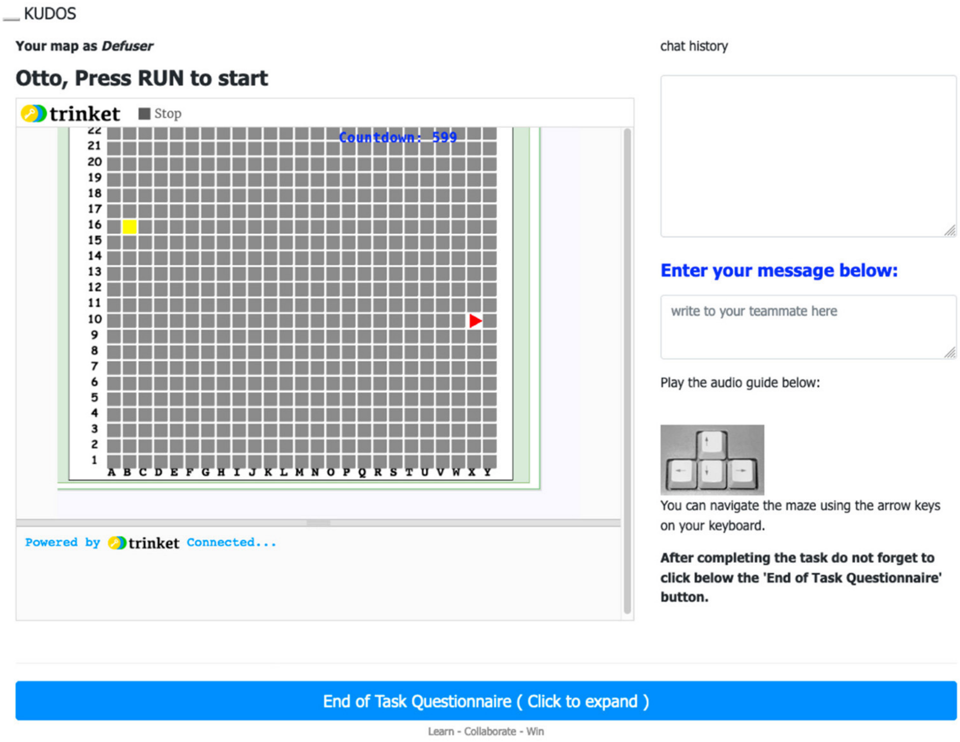
To support the task setting, we designed a custom-made web system. The system pipeline, illustrated in **Figure 1**, was designed according to the following steps:

Step 1: Consent form and registration. Participants registered with a username, AMT IDs (unique identifier needed to reward them at the end of the task), demographic information (gender, age, nationality, and education level), and Big-Five personality traits (**Table 3**). By registering, the participants agreed with the terms of service and gave their informed consent.

Step 2: Introduction and game instructions. After logging in, the "dangerous and challenging world of bomb defusing" (Knuth, 2021), the introductory page offered example screenshots of the two roles, instructions about the gameplay, plus information about the countdown and the end-of-task survey. The short info gave participants a broad idea of the task and focused on the platform functionalities (e.g., chat, game console, manual instructions, etc.).

Step 3: User matching and admin assistance. Participants entered the waiting room (i.e., matchmaking room) and were personally greeted by the system administrator while waiting for their teammates to join. If no other participants were present, they waited until a match would become available. The administrator also served as moderator and user support. The system allocated participants to teams in a first-in-first-out (FIFO) manner. As soon as two participants were present in the matchmaking room, they were placed together and asked to proceed to the main task (after first answering any questions they may have had).

Step 4: Maze challenge and chat box. After matching, participants joined a private virtual room where they could see the maze game and chat to communicate with one another. **Figure 2** shows what the Defuser saw. On the left-hand side, the Defuser saw a blind maze with their position (yellow square) and the bomb (red triangle). They could not see the walls as only the Lead Expert saw them. On the right-hand side, the Defuser saw the chatbox and, below it, a reminder to use the arrow keys to navigate the maze. Upon finishing the task, the blue bar at the bottom of the screen would take them to the final questionnaire. **Figure 3** shows what the Lead Expert saw. The Lead Expert's view of the maze differed from that of the Defuser: they saw only the walls of the maze (gray squares) and the path to the bomb (white sections). The Lead Expert could neither see the Defuser in the maze nor the bomb. Both the Lead Expert and Defuser could see the same countdown and Cartesian coordinates of the maze, as well as the chatbox and the link to the final questionnaire.


KUDOS
Your map as *Defuser*
chat history
trinket
Stop
Countdown: 599
22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
A B C D E F G H I J K L M N O P Q R S T U V W X Y
Powered by trinket Connected...
Enter your message below:
write to your teammate here
End of Task Questionnaire (Click to expand)
Learn - Collaborate - Win

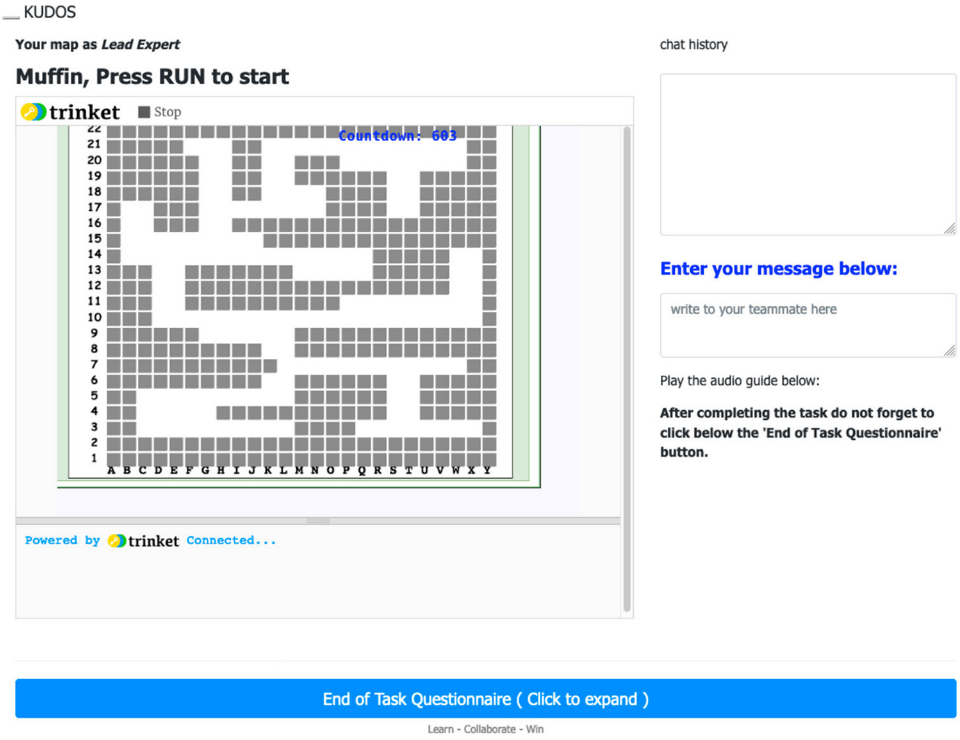
Otto, Press RUN to start

Play the audio guide below:

You can navigate the maze using the arrow keys on your keyboard.

After completing the task do not forget to click below the 'End of Task Questionnaire' button.

FIGURE 2 | Defuser's view of the maze. The maze did not indicate the path to the bomb (red triangle), nor the walls. The participant was prompted to get directions from the Lead Expert through a chatbox (top-right of the screen).


KUDOS
Your map as *Lead Expert*
chat history
trinket
Stop
Countdown: 603
22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
A B C D E F G H I J K L M N O P Q R S T U V W X Y
Powered by trinket Connected...
Enter your message below:
write to your teammate here
End of Task Questionnaire (Click to expand)
Learn - Collaborate - Win

Muffin, Press RUN to start

Play the audio guide below:

After completing the task do not forget to click below the 'End of Task Questionnaire' button.

FIGURE 3 | Lead Expert's view of the maze. The participant could see the map, but did not know where the bomb and the Defuser were placed in the map.

TABLE 2 | Summary of variables.

		Variable	Type	Range
Input	Personality[5]	Extraversion	Interval	2–10
		Agreeableness	Interval	2–10
		Conscientiousness	Interval	2–10
		Emotional stability	Interval	2–10
		Openness to experience	Interval	2–10
	Team Personality (for each trait)	StDev	Ratio	0–5.66
		Min	Interval	2–10
		Max	Interval	2–10
		Mean	Interval	2–10
	Demographics	Gender	Nominal	Male, Female, Other, not-disclosed
		Age group	Ordinal	<20, 20–29, 30–39, 40–49, 50+
		Nationality[6]	Nominal	USA, India, UK, Ireland
		Education level	Ordinal	Less than High School, High School (HS), Some College (SC), College degree (Col), Postgraduate (PG)
	Communication patterns	Uncertainty, Action, Response, Planning, Factual, Non task-related	Ratio	≥0
		Chat length (# Words)	Ratio	≥0
		Chat total (# Posts)	Ratio	≥0
Output	Performance		Nominal	Won, Lost
	Perceived collaboration quality	Performance	Ordinal	1–5
		Cohesion	Ordinal	1–5
		Communication quality	Ordinal	1–5
		Balance	Ordinal	0–2
		Satisfaction	Ordinal	0–2

The Maze module was inspired by the video game "Keep Talking Nobody Explodes" (Knuth, 2021). It consisted of a 25 x 25 grid of squares with one square containing a yellow element (the position of the Defuser), one square containing a red triangle (the position of the bomb), and walls. Neither of the two players had access to all the information of the maze; they needed to cooperate. The Defuser could move inside the maze, by means of the four arrow keys, but they did not know where the walls were. The Lead Expert had the map, but could not navigate the maze. The Defuser's role was to navigate the maze, with the help of the Lead expert, and defuse the bomb in time. Finally, a countdown timer was included, at the end of which the bomb exploded, unless it had been defused. The countdown started the moment both players entered the room. For this specific study, the timer was set to 400 s. After finishing the game, the participants received a validation code to submit to the AMT HIT for getting their base pay and bonus reward (for those teams that completed the challenge successfully). We deliberately excluded aspects of the original video game to reduce the number of variables and increase the controllability of the study environment. We wanted participants to focus on reaching the bomb on time without spreading themselves thin among the multi-modalities present in the original game (e.g., clues, strikes, wires, sequences, etc.). Besides, implementing most features of the original game would have added to the task complexity[4]. Hence, we did not include penalties for the Defuser colliding with a wall. The only penalty—and end of game—was determined by the time running out before reaching the bomb. Furthermore, to ensure task brevity, we considered the bomb defused as soon as the Defuser stepped inside its cell. The simplification of the game has some limitations discussed in Section 6.2.

Step 5: End of task questionnaire. Participants rated the perceived collaboration quality on multiple aspects (see below), and also their abilities.

4. METRICS

We grouped the multilevel approach into two distinct classes referring to input and output variables (**Table 2** provides a summary of all variables, their type and range.). Here the input metrics serve as the independent variables and the output ones as dependent variables.

4.1. Input Variables

4.1.1. Big Five Personality Traits

To acquire a measure of the Big Five traits within the context of large-scale assessment under limited time and resources, we used the Big Five Inventory-10 (BFI-10) (Rammstedt and John, 2007). The inventory consists of ten questions (see **Table 3**).

[4] Also requiring considerably longer instructions and the introduction of manipulation checks to ensure instructions were read which further adds to task complexity.

[5] as the BFI-10 uses 5-point Likert scales one could argue that the data is ordinal, but given a total is calculated per trait we will regard it as interval.

[6] Free text entry, values provided here are those used by participants.

TABLE 3 | BFI-10 instrument used, and its scoring: the trait for which each item was used and whether it was reverse scored (R)[7].

I see myself as someone who …	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	A gree strongly	Trait	Reverted
1. … is reserved	(1)	(2)	(3)	(4)	(5)	Extraversion	R
2. … is generally trusting	(1)	(2)	(3)	(4)	(5)	Agreeableness	
3. … tends to be lazy	(1)	(2)	(3)	(4)	(5)	Conscientiousness	R
4. … is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)	Neuroticism	R
5. … has few artistic interests	(1)	(2)	(3)	(4)	(5)	Openness to Experience	R
6. … is outgoing, sociable	(1)	(2)	(3)	(4)	(5)	Extraversion	
7. … tends to find faults with others	(1)	(2)	(3)	(4)	(5)	Agreeableness	R
8. … does a thorough job	(1)	(2)	(3)	(4)	(5)	Conscientiousness	
9. … gets nervous easily	(1)	(2)	(3)	(4)	(5)	Neuroticism	
10. .. has an active imagination	(1)	(2)	(3)	(4)	(5)	Openness to Experience	

Derived from the shortening of its lengthier predecessor (the Big Five Inventory (BFI-44) Rammstedt and John, 2007), it focuses on the psychometric characteristics of the BFI-44's most representative items and reduces each Big Five dimension to 2 BFI items. The BFI-10 measures the personality traits of Extraversion, Agreeableness, Conscientiousness, Emotional Stability (Neuroticism), and Openness to experience (Rammstedt and John, 2007)[8]. For each trait, the BFI-10 score is calculated as the total score of the two statements associated with that trait, after reversing the score of some statements (see mapping of statements to traits and which statements' scores are reversed in **Table 3**)[9].

4.1.2. Personality Traits of Groups

There is no straightforward process for aggregating metrics such as personality traits for groups. However, the group recommender community has dealt with a similar issue namely the aggregation of group members preferences (Masthoff, 2004) and uses aggregation strategies from Social Choice Theory (Sen, 1986). Senot et al. (2010) distinguishes between (1) majority-based strategies that use the most popular values, (2) consensus-based strategies that consider the profiles of all group members, and (3) borderline strategies that only consider a subset. In our case, majority strategies do not apply given a group size of two. Of the consensus-based strategies, we use Average (which is also the most popular strategy in Group Recommender research). Of the borderline strategies, we use Minimum and Maximum[10],[11]. Minimum is used as one may expect that team performance is strongly affected by the weakest member in the team, in line with the popular saying "a chain is as strong as its weakest link". Maximum is used as one may also expect that a strong member could make up for the weakness in another member (e.g., if one person is highly conscientious, they may entice the team to get the work done in time), particularly when the team is small. Finally, we used Standard Deviation (in line with the Cohesion metric introduced by Odo et al., 2019b), as the literature indicates the impact of diversity within teams[12].

4.1.3. Demographics

Participants provided information about their gender, age group, nationality, and educational background. Socio-demographic measures identify characteristics that often influence the respondent's opinions that could condition one's behavior, culture, and experiences (Lavrakas, 2008). These socio-demographic factors provide further insight into the composition of teams, and what other characteristics—aside from personality traits—influence the collaboration. These socio-demographic factors that make someone distinct can turn into assets for group work. Therefore, by being aware of those characteristics, organizations and hiring bodies can better assemble and coordinate geographically dispersed teams (Muethel et al., 2012).

Multiple studies (Ruef et al., 2003; O'Leary and Mortensen, 2010; Akman et al., 2011) have identified various aspects of the teammates' social backgrounds and demographic characteristics that condition teamwork. For example, members of similar demographic profiles have greater chances to kindle stronger affinity ties (Ruef et al., 2003). Other demographic differences, such as race, sex, age, and nationality, have also been found (Martins and Shalley, 2011) to affect the collective creativity of virtual teams. Age differences condition the creative processes of teams and intensify differences in technical experience (Martins and Shalley, 2011). Differences in nationality have a negative effect by interacting—however indirectly—with the technical experience of the teammates (Martins and Shalley, 2011).

[7] Reverse scored means that a 1 is changed into 5, 2 into 4, 4 into 2, and 5 into 1.

[8] Test-retest correlations suggest acceptable reliability on a Likert scale of 1 (Disagree strongly) to 5 (Agree strongly). As prior studies have shown, the correlations of this instrument with other Big Five instruments, its correlations with self-and peer-ratings, and its associations with socio-demographic variables suggest good validity of the BFI-10 inventory (Rammstedt and John, 2007).

[9] Reversed means that a score of 1 is changed into 5, 2 into 4, 4 into 2, and 5 into 1.

[10] which in the Group Recommender community are called, respectively, Least Misery and Most Pleasure.

[11] Personality traits likely differ on whether a high (or low) trait level positively or negatively impacts team performance. Using both minimum and maximum ensures this is no longer an issue.

[12] For teams of two, the use of standard deviation is equivalent to the use of numerical difference. We opted for standard deviation to build on the work by Odo et al. (2019b) and for generalizability to larger groups.

4.1.4. Communication Patterns

The methodology by Bowers et al. (1998) introduced a new approach to communication analysis prompted by a prior research gap in metrics that missed to analyze the more fine-grained interaction patterns other than simple frequency counts of words. They proposed the implementation of the categories of: (a) **uncertainty** statements, which included direct and indirect questions; (b) **action** statements, which required a particular member to perform a specific action; (c) **acknowledgments**, which were one-bit statements following uncertainty of action statements, such as "yes," "no," "roger"; (d) **responses**, which differed from acknowledgments only in that they conveyed more than one bit of information; (e) **planning** statements; (f) **factual** statements, which verbalized readily observable realities of the environment; and (g) **non task-related** statements. These categories quantified the performance of crews during simulated flight tasks, which improved the make-up of communication sequences analysis.

Based on Bowers et al. (1998) contribution, Davaslioglu et al. (2019) developed the Collective Allostatic Load Measurers system (CALM), which collected, aggregated, and analyzed data from individuals to make assessments on team situation awareness, performance, and resilience. The study used the virtual-reality game “Keep Talking Nobody Explodes” that we too used as inspiration for our experiments. Davaslioglu et al.’s (2019) study demonstrated that some teams exhibited patterns of communication, namely, action-response, uncertainty-response-action, and factual-uncertainty-response-action while working together under high-stress conditions. Acknowledgment statements, for instance, were seen to predominate more amongst high-performing teams, while low-performing teams had higher portions of non-task-related-statements. Similar studies on team communication analysis (Pfaff, 2012; Zijlstra et al., 2012) have identified patterns of communication. Given the proximity of our methodology to the studies of Bowers et al. (1998) and Davaslioglu et al. (2019), we implemented the same communication classes as they did. These communication patterns, or categories, are the following:

- **Uncertainty.** Uncertainty statements comprise questions (either direct or indirect) about the task (e.g., “Where are you at?,” “Where is the bomb?”).
- **Action.** Action statements indicate that one or both of the team members are taking action inside the game, or they are a direction to take action (e.g., “Move two steps down, then one right.” “I am moving to position *x*,” or “Go up for three blocks, then turn right”).
- **Responses.** Response statements can accompany either uncertainty or action statements and suggest that a communication, or feedback loop (e.g., “yes,” “no”), is ongoing.
- **Planning.** Planning statements that give the users a feeling that they are working together toward achieving a common goal. Planning statements can indicate the user’s ability to reassess the situation, clarify the work, or plan the next actions.
- **Factual.** Factual statements are situational and describe the reality, for instance, by giving cues about how the maze looks like from the viewpoint of the Lead Expert, or at which coordinates the bomb is located.
- **Non task-related.** Non-task-related statements are parts of the chats that are categorized as non-related when they do not contribute to the achievement of the goal (e.g., “What is the weather like?”).

TABLE 4 | Example of an annotated chat sequence between a Lead Expert and a Defuser.

Text	Annotation	Role
Okay?	Response	Defuser
Got it?	Response	Lead Expert
I don’t see bomb on my screen, do you know?	Uncertainty	Defuser
I’m the yellow square	Factual	Defuser
czzan’t see bombs	Factual	Lead Expert
where r u?	Uncertainty	Lead Expert
16C	Factual	Defuser
go to 12x	Action	Lead Expert
where should I go?	Uncertainty	Defuser
One step at a time	Planning	Lead Expert
As a lead expert, I request you to guide me	Planning	Lead Expert
Both of us should use the code	Planning	Lead Expert
even I can’t see the bomb	Factual	Lead Expert
there is a triangle on L3	Factual	Defuser
ok	Response	Lead Expert
wait	Action	Lead Expert
can you move? Take turns moving maybe?	Uncertainty	Defuser
follow my steps	Action	Lead Expert
How is your family members?	Non-Related	Defuser

Table 4 illustrates an extract of the annotated chat between the Lead Expert and the Defuser. The patterns were labeled for each participant’s text entry and annotated by two independent evaluators. The inter-rater agreement of the annotation was sufficiently high to be utilized in the study (Cohen’s $\kappa = 0.998$, $p = 0.000$). In addition to counting how often each communication category was used, we also counted the total number of posts made (chat total) and the number of words used (chat length).

4.2. Output Variables

4.2.1. Team Performance

Ancona and Caldwell’s (1992) definition of team performance is the extent to which a team can meet its output targets (e.g., quality, functionality, and reliability of outputs), the expectations of its members, or it’s cost and time goals (Ancona and Caldwell, 1992). For this study, the team performance metric consisted of the binary mapping of the task outcome (winning/losing). The team performance metric has been used as a dependent variable in our functional analysis of the collaboration to illustrate the role of the input factors (personality traits and communication patterns) and allow us to evaluate the constitution of those teams.

4.2.2. Perceived Collaboration Quality

To measure perceived collaboration quality, we use five metrics of team dynamics, which evaluated the participants’ perceptions of their teams.

4.2.2.1. Perceived Performance

The perceived performance metric addresses the question *"How well, in your opinion, did your team perform?."* It was measured on a five-point Likert-scale from *"Very poorly"* (1) to *"Very well"* (5) The perceived performance variable defines the subjective layer of teamwork capability at the given task. The notion has been conceptualized as a multilevel process arising as the teammate engages in their individual and team-level task-work and teamwork processes (Kozlowski and Klein, 2000).

4.2.2.2. Perceived Cohesion

The perceived cohesion metric addresses the question: *"How cohesive was your team?,"* measured using a similar 5-point Likert-scale. Perceived team cohesion, as a fringe term covering social relations, task relations, perceived unity, and emotions (Beal et al., 2003), contributes to our understanding of the emotional dimension of the teams, which is a rather subtle corollary facet of teamwork alongside other subjective measures. The study proposes that group members' perceptions of their cohesion to a particular group are essential in the sense of belonging and feelings of morale (Bollen and Hoyle, 1990). More so, the meta-analysis by Beal et al. (2003) clarifying the construct relation between this particular subjective metric and team performance has denoted a high correlation between these factors across several studies on teams. This work has further established the importance of cohesion (including the subjective measurement) in team performance.

4.2.2.3. Perceived Communication Quality

The perceived communication quality metric addresses the question: *"How well did your team communicate?,"* measured using a similar 5-point Likert-scale. Collecting the perception of the communication quality can help us encode important information about the participant's beliefs toward how a team should function. It can also help disclose the way that the respective individuals engage in communication with the other team members and the way they perceive the communication ties (Cook et al., 2020). Differences in perception might uncover discrepancies between teammates' viewpoints that can lead to the establishment of complex team interventions that intervene at multiple levels of the team formation and interaction processes (Wauben et al., 2011).

4.2.2.4. Perceived Balance

The metric addresses the question: *"Did both members of your team contribute equally in your opinion?"* measured using a 3-point Likert-scale. The variable links with the staging of roles and responsibilities within a team, including how they distribute between teammates and the ways they get carried out against the team's objectives (van de Water et al., 2008). To understand the relevance of the metric within the present study design, remember how entirely different the two roles are and how diametrically determinant they can contribute to teamwork. The top-down allocation of roles was, by itself, not a sufficient guarantee that the teammates' behavior aligned with the given role. By assessing the aspect of perceived balance, through the lenses of the teammates, we could better understand what the participants, and whether it was indeed a balanced act or whether a role was considered more demanding and accountable for the outcome than the other.

4.2.2.5. Satisfaction

The metric addressed the question: *"Would you play with the same teammate again?"* measured using a 3-point Likert-scale. Satisfaction helps predict whether a combination of participants will more likely prefer to work with similar teammates in the future.

5. RESULTS

We divide our results into two themes: 1. performance, and 2. perceived collaboration quality.

1. **Team performance:**
    - **Section 5.1** analyzes the effect of personality at team level [13], comparing winning to losing teams to see if there may be a relationship between personality and performance. It reports the results of a Mann-Whitney U test and perform a regression to investigate the relationship between team traits and the likelihood of a team winning.
    - **Section 5.2** analyzes the communication patterns using a one-way ANOVA to compare winners and losers, but also to compare the differences in behavior between the team roles.
    - **Section 5.3** evaluates the impact on team performance of the participants' socio-demographic characteristics, using Chi-square tests and regression analysis.
2. **Perceived collaboration quality:**
    - **Section 5.4** assesses the relationship between personality traits and perception of collaboration quality, using correlation analysis for the individual traits.
    - **Section 5.5** assesses the relationship between personality traits and perception of collaboration quality, using correlation analysis for the team traits.
    - **Section 5.6** examines whether individual demographic characteristics played any role in people's perception of their collaboration, using one-way ANOVAs.
    - **Section 5.7** analyzes the relationship between the communication patterns and the collaboration quality metrics, also considering the roles of the Defuser and Lead Expert, using correlation analysis.

Given the many factors considered (e.g., considering 5 personality traits with 4 different aggregation metrics for team personality already results in 20 factors) and the many outcome measures, many statistical tests were performed. This may lead to Type I errors. Using Bonferroni corrections[14] to avoid Experiment wide Type I errors would reduce the power of

[13] Team, rather than individual level was used since it is usually the combination and interaction among individuals' personalities that affects the team outcome, as evidenced by multiple studies [e.g., see Gilley et al.'s (2010) comprehensive review].

[14] Less conservative corrections such as Tukey are not possible due to the data often not meeting normality assumptions.

the statistical tests to such an extent that Type II errors would be highly likely and few insights would be gained[15]. We have therefore not applied such corrections (except in *post-hoc* pairwise comparisons). The study is exploratory in nature, and the statistical results presented provide initial insights that lead to hypotheses for follow-on studies.

5.1. Impact of Personality on Team Performance: Minimum Openness May Matter

Since there is no universally accepted way of aggregating team member personality traits into team personality traits, we used multiple, namely the average, minimum, maximum, and standard deviation. Each of these metrics was examined in isolation, as they are not independent. **Table 5** shows the mean (and standard deviation) of these four metrics for the winning and the losing teams. Minimum Openness was significantly better in winning teams (Mann-Whitney U = 485, p = 0.02). There were no other significant results[16].

A binary logistic regression with the minimum metric[17] considered the effects of the teams personality on the likelihood of winning[18]. Given only 16 out of 60 teams won, the basic model only uses a constant with an accuracy of 73.3% (obtained by always predicting the team will lose). The logistic regression model was statistically significant, $\chi^2_{(6)} = 13.60$, $p = 0.034$. The model explained 30% (Nagelkerke R^2) of the variance in winning and correctly classified 77% of cases, including 38% of wins. Increasing minimum Openness and minimum Neuroticism were associated with an increased likelihood of winning [Openness: Exp(B) = 1.52, Wald = 4.61, $p = 0.032$; Neuroticism: Exp(B) = 1.58, Wald = 4.20, $p = 0.041$] .

Our results indicate that in this kind of task (high-pressure, high-demand), minimum Openness to experience seems the most important factor among the Big-5 traits in helping the team to effectively manage the *ad-hoc* collaboration to find a winning solution within a limited time. This means that a crowdsourced, *ad-hoc*, and remote emergency response team will likely be more successful at executing a time-bounded novel task if both collaborators share high levels (minimum) of Openness to experience. The minimum level of this trait indicates that teams with individuals with low Openness are expected to hamper the collaboration regardless of whether the counterpart has very high levels of Openness and this is reasonably determined by the interdependence between roles.

5.2. Impact of Communication Patterns on Team Performance: Action and Response Help Teams Win

Table 6 shows the number of posts per chat category for winners and losers, for winning and losing teams, and for Defusers and Lead Experts. As the role likely affects how participants communicate, we analyzed the communication pattern usage data at the individual level, with an output variable whether these people belonged to winning or losing teams. We analyzed the six chat categories (Uncertainty, Action, Response, Planning, Factual, Non-related), the chat length (in words) and the total number of chat posts between winners and losers using a one-way ANOVA. Winners used significantly more *Action* and *Response* statements [$F_{action}(1{,}118)$ = 4.426, p = 0.038, $F_{response}(1{,}118)$ = 4.983, $p = 0.027$].

A binary logistic regression model to predict whether a participant would win or lose was statistically significant [$\chi^2_{(7)}$ = 14.86, p = 0.038]. The model explained 17% (Nagelkerke R^2) of the variance in winning and correctly classified 78% of cases (25% wins). Increasing the *Action* and *Response* categories was associated with an increased likelihood of winning [Exp(B) = 1.28, Wald = 5.35, p = 0.021; Exp(B) = 1.21, Wald = 3.92, p = 0.048, respectively]. Increasing the chat length was associated with a decreased likelihood of winning [Exp(B) = 0.97, Wald = 4.04, p = 0.044]. These results seem to indicate that participants who gave feedback to one another and focused on discussing which action to take—rather than other types of communication—were able to finish the task and win the game. We also understand that the amount of chat is not a sufficient measure for success in online emergency response team settings since we could not find neither correlation nor causality between these variables.

Lead Experts used the Action category significantly more than Defusers [$F_{action}(1{,}118) = 14.736$, $p < 0.001$] whilst Defusers used the Factual category significantly more [$F_{factual}$ (1, 118) = 5.273, p = 0.023]. The Lead Experts are the ones with the map and would direct the Defusers to the appropriate path to defuse the bomb. Meanwhile, the Defusers may need to tell the Lead Experts where they are. There is a statistically significant difference in the chat categories, with Defusers on winning teams using a significantly higher proportion of Factual messages in their chat than those on losing teams (53 vs. 33%, p = 0.043) and a lower proportion of Uncertainty messages (8 vs. 22%, p = 0.041).

5.3. Impact of Socio-Demographic Characteristics on Performance

Table 7 shows the demographics of winners vs. losers, excluding cases with very low frequency[19]. Pearson Chi-square tests show a significant association between gender and winning [$\chi^2_{(1,N\,=\,119)}$ = 4.78, p = 0.029] and age and winning [$\chi^2_{(3,\,N\,=\,120)}$ = 8.09, p = 0.044]. Men were more likely to win. A binary logistic regression model to predict whether a participant would win or loose based on gender was statistically significant [$\chi^2_{(1)}$ = 5.12,

[15] Additionally, as many measures were not independent, Bonferroni corrections would also have been less appropriate.

[16] Including no impact of Neuroticism or differences of standard deviation.

[17] We only performed the logistic regression with the minimum metric as minimum Openness was the only variable that was significant in the Mann-Whitney test, hence avoiding running multiple tests increasing the chances of Type I error.

[18] Hosmer and Lemeshow test was not significant, thus, the model assumptions were met.

[19] Namely prefer not say for gender, and British and Irish for nationality, all with frequency 1.

TABLE 5 | Mean (Stdev) of standard deviation, average, minimum, and maximum for personality traits for winning and losing teams.

		Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
	StDev	1.06 (0.68)	1.41 (1.46)	1.15 (1.36)	1.50 (1.59)	1.10 (1.00)
Winning	Average	8.13 (1.51)	7.75 (1.53)	5.75 (2.32)	6.94 (1.53)	4.22 (2.33)
Teams	Min	7.38 (1.71)	6.75 (2.24)	4.94 (2.65)	5.88 (2.25)	3.44 (2.42)
	Max	8.87 (1.46)	8.75 (1.34)	6.56 (2.37)	8.00 (1.46)	5.00 (2.45)
	StDev	1.72 (1.36)	1.11 (1.32)	1.66 (1.52)	1.46 (1.25)	1.96 (1.88)
Losing	Average	7.26 (1.60)	8.24 (1.41)	5.01 (1.55)	6.40 (1.26)	3.82 (1.74)
Teams	Min	6.05 (2.22)	7.45 (1.95)	3.84 (1.80)	5.36 (1.79)	2.43 (1.37)
	Max	8.48 (1.42)	9.02 (1.39)	6.18 (1.97)	7.43 (1.25)	5.20 (2.78)

TABLE 6 | Mean (Stdev) of number of times chat categories were used by winners and losers, by winning and losing teams, by Defusers and Lead Experts, and total usage by each.

	Uncertainty	Action	Response	Planning	Factual	Non-related	Total
Winners	2.03 (3.10)	2.91 (4.85)	3.41 (3.77)	0.28 (0.58)	2.34 (2.89)	0.03 (0.18)	11.00 (11.15)
Losers	1.94 (2.30)	1.45 (2.60)	2.14 (2.29)	0.17 (0.49)	2.13 (2.49)	0.52 (2.82)	6.71 (11.00)
Winning teams	4.06 (4.71)	5.81 (6.66)	6.81 (7.08)	0.56 (1.09)	4.69 (4.47)	0.06 (0.25)	22.00 (20.41)
Losing teams	3.89 (3.27)	2.91 (3.67)	4.27 (4.01)	0.34 (0.77)	4.25 (4.21)	1.05 (4.08)	16.70 (11.55)
Defusers	1.62 (2.29)	0.72 (1.29)	2.32 (2.70)	0.27 (0.58)	2.72 (2.87)	0.07 (0.41)	7.70 (6.88)
Lead experts	2.32 (2.72)	2.97 (4.35)	2.63 (2.92)	0.13 (0.43)	1.65 (2.18)	0.72 (3.39)	10.42 (9.14)

p = 0.024]. However, it only explained 6% of the variance in winning and correctly classified 73.1% of cases only by always predicting losing. Being female was associated with a slightly decreased likelihood of winning [Exp(B) = −1.07, Wald = 4.53, p = 0.033]).

We also investigated whether adding gender to the model that uses personality to predict winning would improve the model. A binary logistic regression model to predict whether a participant would win or loose based on gender as well as team personality (in terms of minimum Openness and Neuroticism given the results from Section 5.1) was statistically significant [$\chi^2_{(3)} = 27.97, p < 0.001$]. The model explained 31% (Nagelkerke R^2) of the variance in winning and whilst correctly classifying 78.2% of cases. Being female was associated with a decreased likelihood of winning [Exp(B) = −1.31, Wald = 4.97, p = 0.026]. Similar to our earlier results, increases in minimum Openness and Neuroticism were associated with an increased likelihood of winning [Exp(B) = 0.47, Wald = 11.92, p = 0.001; Exp(B) = 0.52, Wald = 11.94, p = 0.001, respectively]. A similar model without Gender explained only 25% of the variance in winning, and reduced correct classification to 76.5%. Thus, gender mattered but less than personality. When age, nationality or education are added to the binary logistic model instead of gender, they are not significant.

5.4. Impact of Individuals Personality Traits on Perceived Collaboration Quality: Agreeableness May Be Helpful to Cope With Losing

Unfortunately, only 44 out of 120 participants (23 Lead Experts and 21 Defusers) completed the survey at the end of the task, concerning their perception of their team's Cohesion, Performance, Communication, Balance, and Satisfaction. All perceived collaboration metrics were positively correlated (see **Table 8**), overall and for winners. In contrast, for losers the correlations with Satisfaction were not significant (see **Table 8**), and Performance and Balance were also not correlated. So, losers may not always have attributed the bad performance to a poor balance in the team, nor always have been unwilling to keep working with a person even though the collaboration was not going well (according to the other metrics and the fact they lost).

Agreeableness significantly correlated with perceived Performance, Cohesion, and Balance. Neuroticism significantly correlated with only Balance (see **Table 9**). Considering only winners, there were no significant correlations between the personality traits and any metric. In contrast, losers had a significantly positive correlation on Agreeableness with Performance, Cohesion, and Communication. Furthermore, losers had a significant negative correlation on Conscientiousness with Communication. Agreeableness may have helped people to see their loss in a more positive light, making them feel more positively about their teams performance, communication and cohesion[20,21]. We do not know whether being more conscientious made losers feel worse about their teams communication, or whether

[20] This also means that Agreeableness needs to be considered when interpreting indirect measures of team collaboration quality as it may make them a less accurate reflection of actual collaboration.

[21] This seems more likely than that Agreeableness influenced the performance, communication, and cohesion itself, certainly given the lack of correlations for winners.

TABLE 7 | Demographics overall and of winners vs. losers (excluding prefer not to say for gender and nationality) and also for teams that include the same or different genders and nationalities.

	Gender				Nationality				Age				Education			
	Men	**Women**	**Same**	**Differs**	**USA**	**India**	**Same**	**Differs**	**20–29**	**30–39**	**40–49**	**50+**	**HS**	**SC**	**Col**	**PG**
N	78	41	33	27	67	51	33	27	23	56	26	15	9	9	87	15
Winners	33%	15%	30%	22%	19%	35%	27%	26%	22%	36%	27%	0%	11%	33%	28%	27%
Losers	67%	85%	70%	78%	81%	65%	73%	74%	78%	64%	73%	100%	89%	67%	72%	73%

TABLE 8 | Spearman correlations between perceived collaboration quality metrics, $^{**}p < 0.01$, $^{*}p < 0.05$.

		Performance	**Cohesion**	**Communication**	**Balance**	**Satisfaction**
All (*N* = 44)	Performance	1	0.751**	0.593**	0.449**	0.525**
	Cohesion	0.751**	1	0.649**	0.528**	0.502**
	Communication	0.593**	0.649**	1	0.506**	0.508**
	Balance	0.449**	0.528**	0.506**	1	0.389**
	Satisfaction	0.525**	0.502**	0.508**	0.398**	1
Winners (*N* = 24)	Performance	1	0.732**	0.648**	0.486*	0.568**
	Cohesion	0.732**	1	0.725**	0.512*	0.579**
	Communication	0.648**	0.725**	1	0.530**	0.646**
	Balance	0.486*	0.512*	0.530**	1	0.484*
	Satisfaction	0.568**	0.579**	0.646**	0.484*	1
Losers (*N* = 20)	Performance	1	0.734**	0.523*	0.302	0.299
	Cohesion	0.734**	1	0.514*	0.419	0.319
	Communication	0.523*	0.514*	1	0.470*	0.283
	Balance	0.302	0.419	0.470*	1	0.261
	Satisfaction	0.299	0.319	0.283	0.261	1

the team communication was influenced negatively by their Conscientiousness. The lack of a significant correlation for winners points toward the first explanation, with Conscientious people perhaps being more honest in assessing team communication quality.

5.5. Impact of the Teams Personality Traits on Perceived Collaboration Quality: The Positive Role of openness and Surprising Need for Conscientiousness Differences

We determined values for a teams perceived collaboration quality metrics by taking the average of its members, or only one member had provided their ratings by using that members ratings. Average and minimum Openness positively correlated with perceived performance[22] in line with earlier findings that Openness had a positive impact on the likelihood of a team winning. Maximum Agreeableness positively correlated with perceived performance[23], in line with our earlier observations regarding the impact of Agreeableness on individuals opinions.

The most interesting result is the significant positive correlation of all perceived quality metrics with Conscientiousness standard deviation[24,25].

A lower Conscientiousness standard deviation correlated with negative team's feelings. In a dyad, the lowest Conscientiousness standard deviation is when two people work together who are very similar in Conscientiousness. For example, two highly conscientious people or two lowly conscientious people. Two lowly conscientious people working together may not result in a good collaboration. However, two highly conscientious people working together are likely to yield good performance. It seems that the best performance—from the team members' point-of-view—for this particular type of task comes from two people differing in Conscientiousness working together.

5.6. Impact of Socio-Demographic Characteristics on Perceived Collaboration Quality: No Significant Result

Tables 10, **11** show the perceived collaboration quality metrics for the different genders, age groups, nationalities, and education

[22] Spearman correlations average Openness: $r = 0.398$, $p = 0.02$; minimum Openness $r = 0.410$, $p = 0.02$.

[23] Spearman correlation: $r = 0.400$, $p = 0.02$.

[24] Spearman correlations Performance: $r = 0.644$, $p < 0.0001$; Communication quality: $r = 0.492$, $p = 0.003$; Cohesion $r = 0.403$, $p = 0.02$; Balance: $r = 0.448$, $p = 0.008$; Satisfaction: $r = 0.417$, $p = 0.01$.

[25] There was also a significant Spearman correlation for minimum Conscientiousness: $r = -0.423$, $p = 0.01$.

TABLE 9 | Correlations between perceived collaboration quality metrics and personality traits, ***p* < 0.01, **p* < 0.05.

		OPEN	CONS	EXTRO	AGR	NEUR
All (*N* = 44)	Performance	0.062	–0.187	0.044	0.434**	0.106
	Cohesion	0.050	–0.181	–0.088	0.319*	0.160
	Communication	–0.111	–0.256	–0.217	0.221	0.159
	Balance	–0.029	–0.203	–0.196	0.317*	0.318*
	Satisfaction	–0.003	–0.035	–0.074	0.032	–0.031
Winners (*N* = 24)	Performance	0.081	–0.099	0.064	0.289	–0.023
	Cohesion	0.053	–0.148	–0.006	0.241	0.013
	Communication	–0.068	–0.098	–0.239	–0.074	0.044
	Balance	–0.319	–0.302	–0.345	0.354	0.285
	Satisfaction	–0.086	0.144	–0.009	–0.072	–0.098
Losers (*N* = 20)	Performance	0.013	–0.336	0.017	0.761**	0.330
	Cohesion	0.021	–0.226	–0.162	0.456*	0.388
	Communication	–0.178	–0.551*	–0.159	0.547*	0.397
	Balance	0.315	–0.053	0.004	0.338	0.361
	Satisfaction	0.025	–0.233	–0.112	0.242	0.050

TABLE 10 | Mean (standard deviation) of collaboration quality metrics by gender and age, and also for teams that include the same or different genders.

Collaboration	Gender				Age			
	Men (32)	Women (12)	Same (20)	Differs (14)	20–29 (11)	30–39 (25)	40–49 (6)	50+ (2)
Performance	3.75 (1.27)	3.17 (1.53)	3.68 (1.17)	3.21 (1.53)	3.82 (0.87)	3.56 (1.50)	3.50 (1.64)	3.00 (1.41)
Cohesion	3.50 (1.19)	3.00 (1.28)	3.53 (1.09)	3.00 (1.32)	3.55 (1.04)	3.36 (1.22)	2.83 (1.72)	4.00 (0.00)
Communication	3.78 (1.24)	3.25 (1.29)	4.00 (1.06)	2.93 (1.27)	4.27 (0.65)	3.48 (1.33)	3.00 (1.67)	4.00 (0.00)
Balanced	1.03 (0.90)	1.08 (0.67)	1.10 (0.84)	0.89 (0.79)	1.09 (0.83)	1.12 (0.83)	0.33 (0.52)	2.00 (0.00)
Satisfied	1.38 (0.83)	1.08 (0.79)	1.23 (0.83)	1.32 (0.72)	1.27 (0.91)	1.20 (0.82)	1.83 (0.41)	1.00 (1.41)

levels. One-way ANOVAs showed no significant effect of socio-demographic variables on perceived team performance, cohesion, communication, balance, and satisfaction[26]. The averages on all metrics except for balance were a bit higher for men (which would make sense given the men had more often won), but this was not statistically significant, which is not surprising given the high variance and the sample size.

5.7. Impact of Communication Patterns on Perceived Collaboration Quality

We carried out a Spearman correlation test between the communication patterns (the number of occurrences of each communication category for the individual and their team) and the perceived collaboration quality (by individuals[27]).

Satisfaction was positively correlated with the *Factual* category ($r = 0.308$, $p = 0.042$, for both the individual and team), also for Defusers ($r = 0.457$, $p = 0.037$, for the individual), but not Lead Experts. So, members seemed more pleased when their team shared more facts, and Defusers particularly when they shared more facts. Satisfaction was also positively correlated with *Planning* but only for Defusers ($r = 0.437$, $p = 0.047$, for the team). It suggests that Defusers were more pleased when the team planned toward the common goal (i.e., defusing the bomb on time).

Performance was positively correlated with the *Factual* category only for Defusers ($r = 0.504$, $p = 0.020$, for the team). The more cues were shared among the team members the better Defusers seemed to perceive the team performance.

Balance was negatively correlated with the *Uncertainty* category ($r = –0.378$, $p = 0.011$, for the individual), also for Lead Experts ($r = –0.440$, $p = 0.036$; $r = –0.524$, $p = 0.010$, for the individual and team respectively), but not for Defusers. The more questions the Lead Expert asked, and the more questions were asked in the team, the less balanced the Lead Experts seemed to perceive the collaboration.

Finally, *Communication* was positively correlated with the individual *Response* category for Defusers ($r = 0.457$, $p = 0.028$), so the more responsive the Defuser was (e.g., in acknowledging actions they were going to perform), the better they regarded the team communication.

To summarize, several communication categories correlate with perceived collaboration quality, with the role in the team impacting which categories matter. For a good perceived

[26] There was a significant difference for education level on balance, but given the small numbers in all groups.

[27] Given the low number of teams were both members responded, we used the perceived collaboration quality at the individual level only.

TABLE 11 | Mean (standard deviation) of collaboration quality metrics by nationality and education level, and also for teams that include the same or different nationalities.

Collab. Metrics	Nationality				Education Level			
	USA (16)	India (28)	Same (19)	Differs (15)	High Sch. (1)	Some Coll (3)	College (34)	Postgrad. (6)
Performance	3.19 (1.56)	3.82 (1.19)	3.76 (1.25)	3.13 (1.38)	3.00 (0.00)	4.00 (1.00)	3.62 (1.33)	3.33 (1.86)
Cohesion	3.31 (1.40)	3.39 (1.13)	3.53 (1.17)	3.03 (1.22)	3.00 (0.00)	3.33 (0.58)	3.44 (1.16)	3.00 (1.90)
Communication	3.31 (1.49)	3.82 (1.09)	3.68 (1.11)	3.40 (1.44)	2.00 (0.00)	4.67 (0.58)	3.71 (1.12)	3.00 (1.90)
Balanced	1.06 (0.93)	1.04 (0.79)	1.16 (0.78)	0.83 (0.84)	0.00 (0.00)	1.67 (0.58)	1.15 (0.78)	0.33 (0.82)
Satisfied	1.25 (0.86)	1.32 (0.82)	1.40 (0.76)	1.10 (0.81)	1.00 (0.00)	2.00 (0.00)	1.24 (0.82)	1.33 (1.03)

collaboration quality, it seemed important for Defusers to provide facts and neither the team nor the Lead Expert to ask too many questions.

5.8. *Post-hoc* Analysis on Impact of Culture

Given our participants mainly came from the USA and India, one may wonder whether there is an impact of culture. Firstly, whilst there is research to show that personality scales can be generalized across cultures (Rolland, 2002; Rammstedt and John, 2007), the distribution in cultures of personality traits differs. Sometimes therefore statine scores (Thorndike, 1982) are used for personality tests to normalize scores based on participants' country of origin. We did not do this, but did consider how the USA and India differ on personality scores, and whether this difference is visible in our participant sample. **Table 12** shows the personality scores for the USA and India from the literature, and the scores in our sample. In the literature, the main differences between these countries are on Extraversion and Agreeableness. In our sample, there were significant differences in Openness, Extraversion and Agreeableness between the sample from India and the USA[28]. If we had used stanine scoring normalizing based on the country averages from the literature, the difference between the scores in our sample would have been even bigger (given the averages for India where lower than those for the USA in the literature on these traits, and they already are higher than those for the USA in our sample). We conclude that crowd workers recruited through Mechanical Turk do not represent the average person from their countries. This is not surprising, as for example Burnham et al. (2018) found that Mechanical Turkers from the USA are lower in Extraversion than the general USA population (as was also the case in our sample). To be successful on Mechanical Turk, a certain level of conscientiousness is required (as many tasks require a certain success rate on previous tasks). Similarly, one could imagine that coming from India and working on an American platform requires a certain level of Openness to Experience.

There may also be an impact of whether people worked with somebody from their own culture in the task or another culture. We therefore considered whether there was a difference between same nationality teams and teams which differed in nationality on winning the task and on perceptions of collaboration quality (see descriptives in **Tables 7**, **11**, respectively). There was clearly no difference on winning or losing. The perception of collaboration quality seemed slightly better for same nationality teams (with higher means on all measures), but this difference was not statistically significant[29].

6. DISCUSSION, LIMITATIONS, AND FUTURE WORK

6.1. Discussion

In this paper, we explored the impact of personality traits, demographics and communication patterns on a virtual collaborative task under time constraints for crowdsourced dyads. Our study observes how the crowd enacts pair-wise roles under pressure, adjusts its communication via chat, and shares common objectives while executing an artificial, video-game-inspired, cooperative time-bound task. Our goal is to use the knowledge from the observations gathered from the study as the basis for future work on AI-supported crowdsourcing of remote emergency response teams. The main results from our exploration, that will need to be verified in follow-on studies, are as follows:

- **Personality and team performance**: minimum Openness to experience seemed to affect the teams' ability to perform under time pressure. Comparatively, teams with higher minimum Openness levels performed better at the remote cooperative task.
- **Communication and team performance**: Communication patterns seemed to matter for team performance: better-performing crowd teams had more Action/Response statements than non-winning teams.
- **Demographics and team performance**: Gender seemed to influence performance, with men slightly more likely to win, however, gender influenced team performance less than the personality trait Openness to experience (minimum).
- **Personality and perception**: Crowd workers' Agreeableness and Conscientiousness likely shaped their perception of the collaboration. Furthermore, dyads that combined people differing in Conscientiousness were perceived by the participants themselves to perform better.

[28] *Post-hoc* test, Mann-Whitney $U = 811.5$, $U = 611.0$, $U = 933{,}5$ respectively, with $p < 0.001$ (and still significant if Bonferroni corrected).

[29] Perceived performance was significant at $p < 0.05$, but not when Bonferroni correction was applied.

TABLE 12 | Mean and standard deviation of the Big Five personality traits in the literature (Bartram, 2013) and in our sample data.

Data		Openness	Conscientiousness	Extraversion	Agreeableness	Emotional stability
Literature	USA	5.29 (2.05)	5.72 (2.03)	5.84 (2.09)	5.34 (1.97)	5.70 (2.05)
Our sample	USA	6.69 (2.19)	8.34 (1.95)	4.13 (2.02)	5.85 (1.83)	5.88 (2.92)
	India	8.55 (1.56)	7.80 (1.89)	6.71 (1.89)	7.43 (1.74)	6.35 (2.02)

- **Communication and perception**: Communication patterns also seemed to matter for perceived collaboration quality, with the role in the team impacting which categories mattered.

We weigh up these results and connect them with the broader teamwork literature in the coming sections.

6.1.1. Minimum Openness May Impact Teamwork in High-Stress Remote Tasks

Our study demonstrates that the trait of Openness to experience (specifically, its minimum level in a dyadic crowd team) may be a crucial feature for collaboration under pressure and time constraints. This result is novel to the field of team formation since several other studies (Thoms et al., 1996; Barrick et al., 1998; Cogliser et al., 2012; Curşeu et al., 2019) have found that other traits (Conscientiousness first, then Extraversion and Agreeableness) are the most relevant factors affecting team performance. There have been other studies on the effects of personality traits on team performance, such as by O'Neill and Allen (2011) indicating that the trait of Openness is negatively linked with performance *when the team is stable and long-term*, and when it has to perform large analytical tasks such as software engineering. In view of O'Neill and Allen's (2011) study, we read our results as being strongly conditioned by the chosen task type. By highlighting the importance of the trait of Openness, our study helps shed light on the differences that distinguish online *ad-hoc* teams for high-pressure, high-stake tasks, from classical team settings.

Adaptation, as a collateral personality feature of individuals with high Openness to experience, is indeed considered useful in teamwork (Gallivan, 2004), especially in situations of high stress, high-stake and limited time. Moreover, intellectual curiosity with regards to new circumstances is a characteristic observed in people with high Openness to experience (McCrae, 1987); this same trait is closely related to team creativity (Schilpzand et al., 2010). Substantiated by literature (McCrae, 1987; Schilpzand et al., 2010), our results suggest that Openness may act as a more influential factor than task familiarity in determining the success of the team.

6.1.2. Focused Communication Patterns Get the Teams Going

From the results of the analysis of the collaboration, patterns emerge that people who completed the challenge had substantially more Action/Response statements in their chat logs. Thus, they were more effective at communicating with their teammate and promptly came up with clear instructions that helped solve the task on time. Successful participants under pressure used the chat to find a solution right away. Furthermore, winning Defuser predominantly used factual statements. Winning Defusers paid attention to the directives given by their paired teammates (Lead Experts) and responded over the chat by describing where they were at that point in the maze. These results seem to indicate the importance of *focused communication* (with the focus being on efficiency and action clarity), especially when the stakes are high and time-bound. The identification of collaboration patterns has also uncovered tangible clues on how winning individuals intervene during the novel, high-pressure circumstances. Even though communication styles were not communicated explicitly at the start of the task, some participants were more apt at adopting suitable conversational styles as they cooperated and learned from the activity. These findings corroborate other (quasi) longitudinal observations of the long-term impact of risk communication and emergency response measures (Heath and Palenchar, 2000) indicating that citizens are willing to become knowledgeable of emergency response measures and proactively contribute to community relations.

6.1.3. Agreeableness and Conscientiousness Likely Shape the Perception of Collaboration

In our study, highly agreeable people seem to deal better with losing, reflecting more positively on perceived performance, cohesion, and communication. Agreeableness has a social orientation (Bradley et al., 2013) and the trait faceted with trust, altruism, and humility (Matsumoto and Juang, 2016). As highly agreeable people tend to be more sympathetic toward others (Thompson, 2008) and more humble, this may have made them more forgiving toward their teammates and themselves on these aspects. We also found that individuals in teams heterogeneous on Conscientiousness felt better toward the collaboration. Hence, Conscientiousness, at least for high-pressure tasks, is better distributed across teams to improve the perception of teamwork. Making such teams that are heterogeneous in Conscientiousness does not have to be detrimental to actual performance, as shown by our other results as well as Mohammed and Angell (2003). Our result conflicts with that of Gevers and Peeters (2009) who showed that diverse levels of Conscientiousness were negatively linked with teammates' satisfaction. It may be due to the nature of the task since homogeneous high Conscientiousness might have led both the Defuser and the Lead Expert to be overly cautious; however, further studies should investigate the extent of our findings.

6.1.4. Communication Patterns Aligned With Team Roles Matter for the Perception of Collaboration

Communication patterns seemed to matter for the perceived collaboration quality, but this depended heavily on team role. Defusers seemed more satisfied with the collaboration when both themselves and the team used more Factual statements, Lead Experts seemed less satisfied when using Uncertainty statements. These results indicate the importance of team roles and how they are enacted and perceived by teammates. In this instance, the two team roles had distinct and interdependent duties. These reflected the communication patterns that the participants used and preferred (or disliked) above all. In the presence of such distinct team roles, the participants seem to have expected certain communication patterns from their teammates, and these greatly depended on what part of the information they had access to. Defining clear roles is important, as team role clarity improves collaboration (Aritzeta et al., 2005) and communication styles aligned with team roles matter for effective and satisfactory teamwork [as shown in this paper, and in line with (De Vries et al., 2006)]. It may be even more vital in high-pressure tasks with high interdependence.

6.1.5. Gender May Impact Collaboration Though Less Than Personality

Gender seemed to impact team performance, with men slightly more likely to win than women. We considered whether there may have been personality differences. We did not find a statistically significant difference in overall personality traits between genders in this sample. There is some evidence in the literature that there may be a difference in sub-facets of Openness (Weisberg et al., 2011). We also considered whether this is a side effect of the different proportions of men in the sample. More men would result in more teams with men being homogeneous in gender. However, we did not find a significant difference in performance between homogeneous and heterogeneous genders (see **Table 7** for descriptives for same gender teams and teams with different genders). Apesteguia et al. (2012) considered the impact of gender on teamwork in an investment game setting. They argued that a decreased performance in homogeneous female teams is explained by differences in decision making, with women being less aggressive and more focused on social sustainability.

We also considered whether gender homogeneity impacted perceptions of collaboration quality (see **Table 10** for descriptives). There was a significant impact only on Communication (*post-hoc*, Mann Whitney $U = 268$, $p < 0.005$, Bonferroni corrected), with Communication being appreciated more in same gender teams. As there is a big difference between India and the USA in gender equality (USA is 30^{th} (out of 156) in the Global Gender Gap Index (Sharma et al., 2021) compared to India only being 140^{th}), we also considered the impact of gender homogeneity when teams were diverse in nationality. For teams diverse in gender, there was a significant impact of nationality homogeneity on Cohesion and Balance (*post-hoc*, Mann Whitney $U = 28$, $p < 0.05$, Bonferroni corrected) and similar trends for Communication and Performance ($p = 0.1$ after Bonferroni correction), with all being perceived better for same nationality teams. We considered whether the impact of gender on winning we found may be partially due to women being more likely to have been in diverse gender teams, and collaboration issues having occurred in such teams when the teams were mixed in nationality. However, this was not supported by the data. Further studies are needed to investigate possible cultural factors and their interaction with gender homogeneity. However, given the impact gender may have, gender diversity in teams should be encouraged (Díaz-García et al., 2013).

6.2. Limitations

6.2.1. Exploratory Study

As explained above, the study performed was exploratory in nature. Follow-on studies are needed to confirm the results found. The findings from our study can provide the hypotheses for such studies.

6.2.2. Matchmaking System

One of the primary limitations of this study comes from the matchmaking part of the system. We paired participants following a simple first-in-first-out queuing fashion and did not consider user features. This study design choice matched the micro-tasking nature of crowdsourcing and its asynchronous environment, characteristics typical to platforms like Amazon Mechanical Turk. Random matching proved to be an effective solution to the problem of pairing virtual users into *ad-hoc* teams fast and based on availability, and for this reason easily applicable in emergencies. However, this matching limited the control over team formation, rendering the present study observational. For future studies, we plan to test other types of matchmaking criteria. For example, using heuristic algorithms similar to Irvin's Stable Roommate Problem (Irving, 1985) that would assist the matchmaking process according to pre-defined criteria. Other matching systems, such as AI (machine learning and features extraction), could also be used as baselines.

6.2.3. Metrics and Sample

Another limitation of this study is the one associated with the dataset generated from the user outputs and their willingness to give away credible information on their personality traits, demographic data, and experience in the game. We plan to strengthen this area of the research by implementing additional types of secondary data collection systems, such as behavioral, contextual, and sensor data, to help validate and enrich the information gathered about the participants. Different user groups (e.g., students, remote developers, and incident response volunteers) should partake in future studies.

Additionally, our study design did not implement exclusion criteria such as required English proficiency levels nor relied upon pre-screening to filter crowd workers on the basis of their reputation and/or a number of successful HITs. Varying levels of English may have impacted the results. However, most participants reported having completed a College education and the education language at College in all participants' countries (USA, India, UK, Ireland) is English, so we have some confidence that the English level was sufficient not to inhibit communication. We also did not notice clear communication issues due to language in the chats. Nevertheless, future studies will include a test to ensure an appropriate English proficiency level. The absence of pre-screening on English also has a positive aspect, as means our study can be generalized to emergency crises where

English is not necessarily the native language whilst still being used for basic virtual communication via chat.

Finally, our sample consisted of predominately male, American, and Indian AMT workers. The sample used for the results likely impacted participants' collaboration and performance. Although we accounted for some of these socio-demographic characteristics (of which gender was significant), we acknowledge the limitations of the dataset derived from the AMT sample. Other types of remote crowd workers from other platforms should experiment with the tool to test for the generalisability of the findings to other portions of the population.

6.2.4. Task, Timer, and Features

The results gathered from the experiments on a single task provide a limited range of conclusions and levels of abstraction to other domains unless other high-stress scenarios could be tested and compared. We plan to implement several types of high-stress tasks. For instance, real-time translation or visual puzzle games would generate more diverse data. They would also quantify the extent to which the choice of task design impacts team collaboration. Another limitation is the lack of manipulation checks for the perceived realism and urgency of the task. It is possible that those workers who did not approach the task seriously might have behaved differently in situations of authentic danger and gravity. Future work should apply similar methodologies and observations to real-life remote emergency situations to be able to test the generalizability of our findings[30]. As part of the development stage, we ran several pilot studies to improve the initial task design and make the instructions clear and understandable for the participating crowd workers.

In the process, we omitted multiple elements present in the original version of the module. We tested different countdowns during the pre-study phase with real users. We settled for a timelimit of 400s as it allowed participants to familiarize themselves with the task interface, chat with one another, and execute the task. Time limits can still be the subject of further testing to evaluate the user's reaction times. We deliberately excluded some of the original elements of the maze module from the video game (i.e., the count of strikes or penalty points for hitting the invisible blocks when crossing the walls, the view of the multiple mazes from the Lead Expert manual, etc.). Tweaking in-game parameters will help uncovering differences in behavior and collaboration that we could not identify by running a single study design. In our experiments, the maze's walls were made invisible to the Defuser while still detectable through object collision. In future studies, and as part of the task improvements, we aim to bring back some of the original features and to assess their significance.

6.3. Implications and Future Work

6.3.1. AI Support for Team Formation in Emergency Response

There has been growing research on AI supported team formation, where AI programs allocate workers or learners to teams (Lykourentzou et al., 2013; Odo et al., 2019a). Clearly, the task impacts what team attributes matter for good actual and perceived performance and collaboration. For the emergency task studied in this paper, our primary finding concerns the importance of the trait of Openness to Experience (minimum). When developing an AI group formation system, this can be incorporated (e.g., in the criteria used for automated team formation), ensuring emergency response team have high minimum Openness to Experience, and diverting crowd workers with low Openness to more suitable tasks. Pre-screening and selection procedures are not new to disaster management, but our findings indicate that certain personality traits affect emergency teamwork, and this goes beyond the more common filtering criteria used such as reputation and trust (Javaid et al., 2013). More so, previous research on the effects of personality traits in teamwork did not consider the impact of the task type under stress (Thoms et al., 1996; Barrick et al., 1998; Cogliser et al., 2012; Curşeu et al., 2019), particularly in cases of emergency response. The sample of crowd workers used in this study helped us understand how pairs of non-familiar and dispersed users act together when presented with an unseen challenge. By utilizing AI to infer the crowd's attributes through their interactions, intelligent systems can learn to adjust to their needs and capabilities in times of emergency and suggest collaborators for a better fit.

The results from this specific approach are beneficial to the crowdsourcing and online work fields that are becoming ever so relevant due to recent and significant changes in the way we live and work. In the Ukrainian conflict of 2022, volunteers of remote rescue operations based in the USA allocated buses to civilians making requests for help online and helping save countless lives (Mark et al., 2022). By remote communication and real-life GPS updates, citizens from far away aided the evacuation of many citizens by identifying grounds hit by shelling and bombing. Following tragic examples like this one, researchers and industry can weigh the power of AI to aid the team formation process of remote emergency crowd teams and assist with organizing rescue units during high-stress, life-threatening situations.

6.3.2. Conversational AI Support for Remote Emergency Response Teams

The analysis of the communication patterns clearly indicated that not all teams focused on the task execution correctly since some adopted less-than-optimal communication strategies. Our results provide insights into which communication acts may be important which can be used by an AI system to monitor and moderate remote collaboration and intervene when needed. With the implementation of machine learning models, future crowdsourcing tools specialized in emergency response can augment the chat functionality by deploying conversational AI (Battineni et al., 2020) (as an example) moderating users' communication patterns. With the stark improvements in Natural Language Generation, Understanding, and Processing, and the increasingly reduced costs of production thanks to open-source software community (Adamopoulou and Moussiades, 2020), most forms of crowdsourced self-organized teams (e.g., neighborhood watch Bakker et al., 2012) could themselves

[30] However, there are clear ethical issues with this.

incorporate, maintain, and improve machine learning models for emergency response conversational AI initially trained on annotations and knowledge such as the one we present.

We note that personality traits seemed to affect the perception of the collaboration. Although system evaluations usually pursue metrics similar to ours (e.g., effectiveness, efficiency, and reliability), team performance is only part of the equation. While a team can successfully reach a goal on time, the perception of teamwork is not always directly proportional to that outcome. What individuals think, interpret, and how they respond to changes can be conditioned by personality factors. In this study, we observe the interaction between personality and communication patterns. With defined team roles and interdependency, people with certain personality traits are likely to expect from others certain communication styles. Further, personality seems to have determined the propensity for more or less rigor and clarity in the communication. Considering the numerous variables at play and the increased reliance on crowdsourcing for rescue operations and emergency response (Marc Cieslak, 2022), we advocate for the development of adaptive and personalized intelligent systems. AI-aided emergency response can provide support and knowledge to teams according to the individual and group needs to alleviate stress and improve community participation. Emotional support could be tailored to the individuals and made accessible and private in critical emergency settings addressing the lack of sensemaking and trust emerging from periods of stress, trauma, and danger.

7. CONCLUSION

In this study, 60 crowd dyads collaborated in a high-pressure, computer-mediated task. The study required them to play complementary roles in a time-bounded critical scenario. We explored the possible impact of the participants' personality, socio-demographic factors, and communication patterns on team performance and perceived collaboration quality. Results from our exploratory study suggest that teams scoring high on the personality trait of Openness (meaning that the minimum Openness of winning teams was higher than in the losing teams) performed overall better in the execution of this high-pressure task. The analysis of the team communication patterns suggest that teams communicating more through action-response loops were more likely to win the game. Different levels of Agreeableness and Conscientiousness likely shaped the perception of collaboration with highly agreeable people coping better with losing. Teams heterogeneous on Conscientiousness seemed to feel better about the teamwork. Communication patterns seemed to matter for the perceived collaboration quality, but this was highly role-dependent, showing that communication styles aligned with team roles matter for effective and satisfactory teamwork. We can learn from these exploratory results that the perception of the collaboration may differ depending on personality traits and the communication patterns shared among remote teammates. So, intelligent crowdsourcing-aided emergency response technology may need to consider individuals' viewpoints and provide adequate support for the crowd needs. Our findings support future research on computer-based collaboration under pressure. It shows ways to tailor the development of AI as accessible support in crowdsourcing emergency response aiding with team formation, conversational support, and adaptation. Future work will confirm the findings and evaluate other types of high-stress tasks, time limits, and parameters for team formation to advance the findings presented here.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Abdul-Razik, M. S., Kaity, A. M., Banafaa, N. S., and El-Hady, G. W. (2021). Disaster response in a civil war: lessons on local hospitals capacity. the case of yemen. *Int. J. Healthcare Manag.* 14, 99–106. doi: 10.1080/20479700.2019.1616386

Abu-Elkheir, M., Hassanein, H. S., and Oteafy, S. M. (2016). "Enhancing emergency response systems through leveraging crowdsensing and heterogeneous data," in *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)* (Paphos: IEEE), 188–193.

Adamopoulou, E., and Moussiades, L. (2020). "An overview of chatbot technology," in *IFIP International Conference on Artificial Intelligence Applications and* Innovations (Crete: Springer), 373–383.

Akman, I., Misra, S., and Altindag, T. (2011). The impact of cognitive and socio-demographic factors at meetings during software development process. *Tehnički vjesnik* 18, 51–56. Available online at: https://hrcak.srce.hr/65924

Alexander, D. (2003). Towards the development of standards in emergency management training and education. *Disaster Prevent. Manag.* 12, 113–123. doi: 10.1108/09653560310474223

Allen, D. W. (2011). *Getting to Know ArcGIS: Modelbuilder, Vol. 380*. Seattle, WA: Esri Press Redlands.

Ancona, D. G., and Caldwell, D. F. (1992). Demography and design: Predictors of new product team performance. *Organ. Sci.* 3, 321–341. doi: 10.1287/orsc.3.3.321

Anderson, M. (2020). *How the Help Desk Can Support the Security Team* (Ph.D. thesis), Utica College.

Apesteguia, J., Azmat, G., and Iriberri, N. (2012). The impact of gender composition on team performance and decision making: evidence from the field. *Manag. Sci*. 58, 78–93. doi: 10.1287/mnsc.1110.1348

Archibold, R. C. (2003). *A Nation at war: The New Organizations; Papers Debate Use of Teams to React Quickly to Terror*. New York Time.

Aritzeta, A., Ayestaran, S., and Swailes, S. (2005). Team role preference and conflict management styles. *Int. J. Conflict Manag*. 16, 157–182. doi: 10.1108/eb022927

Bakker, J., Denters, B., Oude Vrielink, M., and Klok, P.-J. (2012). Citizens initiatives: How local governments fill their facilitative role. *Local Govern. Stud*. 38, 395–414. doi: 10.1080/03003930.2012.698240

Baldwin, S., and Woods, P. A. (1994). Case management and needs assessment: some issues of concern for the caring professions. *J. Mental Health* 3, 311–322. doi: 10.3109/09638239408997941

Barrick, M. R., Stewart, G. L., Neubert, M. J., and Mount, M. K. (1998). Relating member ability and personality to work-team processes and team *Effectiveness* 83, 377–391. doi: 10.1037/0021-9010.83.3.377

Bartram, D. (2013). Scalar equivalence of opq32: Big five profiles of 31 countries. *J. Cross Cult. Psychol*. 44, 61–83. doi: 10.1177/0022022111430258

Battineni, G., Chintalapudi, N., and Amenta, F. (2020). Ai chatbot design during an epidemic like the novel coronavirus. *Healthcare* 8, 154. doi: 10.3390/healthcare8020154

Beal, D. J., Cohen, R. R., Burke, M. J., and McLendon, C. L. (2003). Cohesion and performance in groups: a meta-analytic clarification of construct relations. *J. Appl. Psychol*. 88, 989. doi: 10.1037/0021-9010.88.6.989

Belbin, R. M. (2012). *Team Roles at Work*. Oxon: Routledge.

Bell, S. T., Fisher, D. M., Brown, S. G., and Mann, K. E. (2018). An approach for conducting actionable research with extreme teams. *J. Manag*.. 44, 2740–2765. doi: 10.1177/0149206316653805

Bergen, M. (2020). *Google outage reignites worries about smart home without backups*. Available online at: https://www.bloomberg.com/news/newsletters/2020-12-16/google-outage-reignites-worries-about-smart-home-without-backups

Bjerge, B., Clark, N., Fisker, P., and Raju, E. (2016). Technology and information sharing in disaster relief. *PLoS ONE* 11, e0161783. doi: 10.1371/journal.pone.0161783

Bollen, K. A., and Hoyle, R. H. (1990). Perceived cohesion: a conceptual and empirical examination. *Soc. Forces* 69, 479–504. doi: 10.2307/2579670

Borman, W. C., Rosse, R. L., and Abrahams, N. M. (1980). An empirical construct validity approach to studying predictor-job performance links. *J. Appl. Psychol*. 65, 662. doi: 10.1037/0021-9010.65.6.662

Bowers, C. A., Jentsch, F., Salas, E., and Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Hum. Factors* 40, 672–679. doi: 10.1518/001872098779649265

Boyd, R., and Brown, T. (2005). Pilot study of myers briggs type indicator personality profiling in emergency department senior medical staff. *Emergency Med. Aust*. 17, 200–203. doi: 10.1111/j.1742-6723.2005.00723.x

Boyle, G. J. (1995). Myers-briggs type indicator (mbti): some psychometric limitations. *Aust. Psychol*. 30, 71–74. doi: 10.1111/j.1742-9544.1995.tb01750.x

Bradley, B. H., Baur, J. E., Banford, C. G., and Postlethwaite, B. E. (2013). Team players and collective performance: how agreeableness affects team performance over time. *Small Group Res*. 44, 680–711. doi: 10.1177/1046496413507609

Brennan, M. A., and Flint, C. G. (2007). Uncovering the hidden dimensions of rural disaster mitigation: capacity building through community emergency response teams. *J. Rural Soc. Sci*. 22, 7. Available online at: https://egrove.olemiss.edu/jrss/vol22/iss2/7

Briggs, S. M. (2005). Disaster management teams. *Curr. Opin. Crit. Care* 11, 585–589. doi: 10.1097/01.ccx.00001186916.92757.ab

Burnham, M. J., Le, Y. K., and Piedmont, R. L. (2018). Who is mturk? personal characteristics and sample consistency of these online workers. *Mental Health Religion Cult*. 21, 934–944. doi: 10.1080/13674676.2018.1486394

Capraro, R. M., and Capraro, M. M. (2002). Myers-briggs type indicator score reliability across: studies a meta-analytic reliability generalization study. *Educ. Psychol. Meas*. 62, 590–602. doi: 10.1177/0013164402062004004

Careem, M., De Silva, C., De Silva, R., Raschid, L., and Weerawarana, S. (2006). "Sahana: overview of a disaster management system," in *2006 International Conference on Information and Automation* (Colombo: IEEE), 361–366.

Carver, L., and Turoff, M. (2007). Human-computer interaction: the human and computer as a team in emergency management information systems. *Commun. ACM* 50, 33–38. doi: 10.1145/1226736.1226761

Chau, M. M. (2020). Rapid response to a tree seed conservation challenge in hawai ʻi through crowdsourcing, citizen science, and community engagement. *J. Sustain. Forestry* 1–19. doi: 10.1080/10549811.2020.1791186. [Epub ahead of print].

Chen, L., and Miller-Hooks, E. (2012). Optimal team deployment in urban search and rescue. *Transport. Res. B Methodol*. 46, 984–999. doi: 10.1016/j.trb.2012.03.004

Chen, R., Sharman, R., Rao, H. R., and Upadhyaya, S. J. (2008). Coordination in emergency response management. *Commun. ACM* 51, 66–73. doi: 10.1145/1342327.1342340

Cogliser, C. C., Gardner, W. L., Gavin, M. B., and Broberg, J. C. (2012). Big five personality factors and leader emergence in virtual teams: relationships with team trustworthiness, member performance contributions, and team performance. *Group Organ. Manag*. 37, 752–784. doi: 10.1177/1059601112464266

Colovic, A., Caloffi, A., and Rossi, F. (2022). Crowdsourcing and covid-19: how public administrations mobilize crowds to find solutions to problems posed by the pandemic. *Public Adm Rev*. doi: 10.1111/puar.13489. [Epub ahead of print].

Cook, A., Zill, A., and Meyer, B. (2020). Perceiving leadership structures in teams: Effects of cognitive schemas and perceived communication. *Small Group Res*. 25, 251–287. doi: 10.1177/1046496420950480

Curşeu, P. L., Ilies, R., Vîrgă, D., Maricuţoiu, L., and Sava, F. A. (2019). Personality characteristics that are valued in teams: not always "more is better"? *Int. J. Psychol*. 54, 638–649. doi: 10.1002/ijop.12511

Davaslioglu, K., Pokorny, B., Sagduyu, Y. E., Molintas, H., Soltani, S., Grossman, R., et al. (2019). "Measuring the collective allostatic load," in *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)* (Las Vegas, NV: IEEE), 93–99.

De Vries, R. E., Van den Hooff, B., and de Ridder, J. A. (2006). Explaining knowledge sharing: the role of team communication styles, job satisfaction, and performance beliefs. *Commun. Res*. 33, 115–135. doi: 10.1177/0093650205285366

De Wit, F. R., Greer, L. L., and Jehn, K. A. (2012). The paradox of intragroup conflict: a meta-analysis. *J. Appl. Psychol*. 97, 360. doi: 10.1037/a0024844

Díaz-García, C., González-Moreno, A., and Jose Saez-Martinez, F. (2013). Gender diversity within r&d teams: its impact on radicalness of innovation. *Innovation* 15, 149–160. doi: 10.5172/impp.2013.15.2.149

Difallah, D., Filatova, E., and Ipeirotis, P. (2018). "Demographics and dynamics of mechanical turk workers," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. (Marina Del Rey, CA: ACM), 135–143.

Dilmaghani, R. B., and Rao, R. R. (2006). "On designing communication networks for emergency situations," in *2006 IEEE International Symposium on Technology and Society* (Queens, NY: IEEE), 1–8.

Driskell, T., Burke, S., Driskell, J., Salas, E., and Neuberger, L. (2014). Steeling the team: assessing individual and team functioning "at a distance.". *Military Psychol*. 29, 12–18.

Driskell, T., Salas, E., and Driskell, J. E. (2018). Teams in extreme environments: Alterations in team development and teamwork. *Hum. Resour. Manag. Rev*. 28, 434–449. doi: 10.1016/j.hrmr.2017.01.002

Ducao, A. A. B. (2013). *OpenIR [Open Infrared*]: *enhancing environmental monitoring through accessible remote sensing, in Indonesia and beyond* (Ph.D. thesis), Massachusetts Institute of Technology.

Ellis, A. P., Bell, B. S., Ployhart, R. E., Hollenbeck, J. R., and Ilgen, D. R. (2005). An evaluation of generic teamwork skills training with action teams: effects on cognitive and skill-based outcomes. *Pers. Psychol*. 58, 641–672. doi: 10.1111/j.1744-6570.2005.00617.x

Elsafoury, F. (2020). "Teargas, water cannons and twitter: a case study on detecting protest repression events in turkey 2013," in *Text2Story@ ECIR*. (Lisbon), 5–13.

Ernst, C., Mladenow, A., and Strauss, C. (2017). Collaboration and crowdsourcing in emergency management. *Int. J. Pervasive Comput. Commun.* 13, 176–193. doi: 10.1108/IJPCC-03-2017-0026

Etheridge, J. C., Moyal-Smith, R., Sonnay, Y., Brindle, M. E., Yong, T. T., Tan, H. K., et al. (2022). Non-technical skills in surgery during the covid-19 pandemic: an observational study. *Int. J. Surg.* 98, 106210. doi: 10.1016/j.ijsu.2021.106210

Farkas, J., and Neumayer, C. (2017). 'stop fake hate profiles on facebook: Challenges for crowdsourced activism on social media. *First Monday*. 22. doi: 10.5210/fm.v22i9.8042

Flin, R., and Slaven, G. (1996). Personality and emergency command ability. *Disaster Prevent. Manag.* 5, 40–46. doi: 10.1108/09653569610109550

Floch, J., Angermann, M., Jennings, E., and Roddy, M. (2012). "Exploring cooperating smart spaces for efficient collaboration in disaster management," in *ISCRAM*. (Vancouver, BC).

Foushee, H. C. (1984). Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. *Am. Psychol.* 39, 885. doi: 10.1037/0003-066X.39.8.885

Friede, A. M. (2022). In defence of the baltic sea region:(non-) allied policy responses to the exogenous shock of the ukraine crisis. *Eur. Security* 1–23. doi: 10.1080/09662839.2022.2031990. [Epub ahead of print].

Galbraith, J. R., and Lawler, E. E. (1993). *Organizing for the Future: The New Logic for Managing Complex Organizations*. Durham: Jossey-Bass Inc. Publisher.

Gallivan, M. J. (2004). Examining it professionals' adaptation to technological change: The influence of gender and personal attributes. *ACM SIGMIS Database* 35, 28–49. doi: 10.1145/1017114.1017119

Gardner, H. K. (2012). Performance pressure as a double-edged sword: enhancing team motivation but undermining the use of team knowledge. *Administrat. Sci. Q.* 57, 1–46. doi: 10.1177/0001839212446454

Gay-Antaki, M. (2021). Stories from the ipcc: an essay on climate science in fourteen questions. *Glob. Environ. Change* 71, 102384. doi: 10.1016/j.gloenvcha.2021.102384

Gersick, C. J., and Hackman, J. R. (1990). Habitual routines in task-performing groups. *Organ. Behav. Hum. Decis. Process.* 47, 65–97. doi: 10.1016/0749-5978(90)90047-D

Gevers, J. M. P., and Peeters, M. A. G. (2009). A pleasure working together? the effects of dissimilarity in team member conscientiousness on team temporal processes and individual satisfaction. *J. Organ. Behav.* 30, 379–400. doi: 10.1002/job.544

Gilley, J. W., Morris, M. L., Waite, A. M., Coates, T., and Veliquette, A. (2010). Integrated theoretical model for building effective teams. *Adv. Dev. Hum. Resour.* 12, 7–28. doi: 10.1177/1523422310365309

Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *J. Pers. Soc. Psychol.* 59, 1216. doi: 10.1037/0022-3514.59.6.1216

Hackman, J. R. (1990). *Groups that work and those that don't*. Number E10 H123. Jossey-Bass.

Hackman, J. R., and Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: a review and proposed integration. *Adv. Exp. Soc. Psychol.* 8, 45–99. doi: 10.1016/S0065-2601(08)60248-8

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). "A data-driven analysis of workers' earnings on amazon mechanical turk," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. (Montreal, QC), 1–14.

Harrison, A. A., and Connors, M. M. (1984). Groups in exotic environments. *Adv. Exp. Soc. Psychol.* 18, 49–87. doi: 10.1016/S0065-2601(08)60142-2

Hart, J. D., Piumsomboon, T., Lawrence, L., Lee, G. A., Smith, R. T., and Billinghurst, M. (2018). "Demonstrating emotion sharing and augmentation in cooperative virtual reality games," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. (Melbourne, VIC), 117–120.

Heath, C., and Luff, P. (1992). Collaboration and controlcrisis management and multimedia technology in london underground line control rooms. *Comput. Support. Cooperative Work* 1, 69–94. doi: 10.1007/BF00752451

Heath, R. L., and Palenchar, M. (2000). Community relations and risk communication: a longitudinal study of the impact of emergency response messages. *J. Public Relat. Res.* 12, 131–161. doi: 10.1207/S1532754XJPRR1202_1

Heinzelman, J., and Waters, C. (2010). *Crowdsourcing Crisis Information in Disaster-Affected Haiti*. Washington, DC: U.S. Institute of Peace.

Helmreich, R. L. (1967). *Prolonged stress in sealab ii: A field study of individual and group reactions*. Technical report, Yale University New Haven Ct.

Hemphill, L., and Roback, A. J. (2014). "Tweet acts: how constituents lobby congress via twitter," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. (Baltimore, MD: ACM), 1200–1210.

Herstein, J. J., Schwedhelm, M. M., Vasa, A., Biddinger, P. D., and Hewlett, A. L. (2021). Emergency preparedness: What is the future? *Antimicrob. Stewardship Healthcare Epidemiol.* 1, e29. doi: 10.1017/ash.2021.190

Highhouse, S., Wang, Y., and Zhang, D. C. (2022). Is risk propensity unique from the big five factors of personality? a meta-analytic investigation. *J. Res. Pers.* 98, 104206. doi: 10.1016/j.jrp.2022.104206

Hossain, A., Mirza, F., Naeem, M. A., and Gutierrez, J. (2017). "A crowd sourced framework for neighbour assisted medical emergency system," in *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)* (Melbourne, VIC: IEEE), 1–6.

Hughes, J. A., Randall, D., and Shapiro, D. (1992). "Faltering from ethnography to design," in *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work*. (Toronto, ON: ACM), 115–122.

Ikizer, G., Kowal, M., Aldemir, İ. D., Jeftić, A., Memisoglu-Sanli, A., Najmussaqib, A., et al. (2022). Big five traits predict stress and loneliness during the covid-19 pandemic: evidence for the role of neuroticism. *Pers. Individ. Differ.* 190, 111531. doi: 10.1016/j.paid.2022.111531

Irving, R. W. (1985). An efficient algorithm for the "stable roommates" problem. *J. Algorithms* 6, 577–595. doi: 10.1016/0196-6774(85)90033-1

Javaid, S., Majeed, A., and Afzal, H. (2013). "A reputation management system for efficient selection of disaster management team," in *2013 15th International Conference on Advanced Communications Technology (ICACT)* (PyeongChang: IEEE), 829–834.

Kamphuis, W., Gaillard, A. W., and Vogelaar, A. L. (2011). The effects of physical threat on team processes during complex task performance. *Small Group Res.* 42, 700–729. doi: 10.1177/1046496411407522

Kanki, B. G., Folk, V. G., and Irwin, C. M. (1991). Communication variations and aircrew performance. *Int. J. Aviat. Psychol.* 1, 149–162. doi: 10.1207/s15327108ijap0102_5

Kanki, B. G., Lozito, S., and Foushee, H. C. (1989). "Communication indices of crew coordination." in *Aviation, Space, and Environmental Medicine.* (Aerospace Medical Assn).

Kawatsuma, S., Fukushima, M., and Okada, T. (2012). Emergency response by robots to fukushima-daiichi accident: summary and lessons learned. *Ind. Robot*. 39, 428–435. doi: 10.1108/01439911211249715

Kennedy, B., Curtis, K., and Waters, D. (2014). The personality of emergency nurses: is it unique? *Aust. Emerg. Nurs. J.* 17, 139–145. doi: 10.1016/j.aenj.2014.07.002

Kichuk, S. L., and Wiesner, W. H. (1997). The big five personality factors and team performance: implications for selecting successful product design teams. *J. Eng. Technol. Manag.* 14, 195–221. doi: 10.1016/S0923-4748(97)00010-6

Kincaid, J. P., Donovan, J., and Pettitt, B. (2003). Simulation techniques for training emergency response. *Int. J. Emerg. Manag.* 1, 238–246. doi: 10.1504/IJEM.2003.003300

King, G. (2002). Crisis management team effectiveness: a closer examination. *J. Bus. Ethics* 41, 235–249. doi: 10.1023/A:1021200514323

Knuth, D. (2021). Steelcrategames. Available online at: https://twitter.com/SteelCrateGames

Kozlowski, S. W., and Klein, K. J. (2000). *A Multilevel Approach to Theory and Research in Organizations: Contextual, Temporal, and Emergent Processes*. Jossey-Bass.

Kretzschmar, M. E., Ashby, B., Fearon, E., Overton, C. E., Panovska-Griffiths, J., Pellis, L., et al. (2022). Challenges for modelling interventions for future pandemics. *Epidemics* 38, 100546. doi: 10.1016/j.epidem.2022.100546

Krumm, S., Kanthak, J., Hartmann, K., and Hertel, G. (2016). What does it take to be a virtual team player? the knowledge, skills, abilities, and other characteristics required in virtual teams. *Hum. Perform.* 29, 123–142. doi: 10.1080/08959285.2016.1154061

Lacoursiere, R. B. (1980). *The Life Cycle of Groups: Group Developmental Stage Theory*. New York, NY: Human Sciences Press.

Landgren, J., and Nulden, U. (2007). "A study of emergency response work: patterns of mobile phone interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (San Jose, CA), 1323–1332.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Sage Publications.

Leach, L. S., and Mayo, A. M. (2013). Rapid response teams: qualitative analysis of their effectiveness. *Am. J. Crit. Care* 22, 198–210. doi: 10.4037/ajcc2013990

Leach, M., MacGregor, H., Ripoll, S., Scoones, I., and Wilkinson, A. (2022). Rethinking disease preparedness: incertitude and the politics of knowledge. *Crit. Public Health* 32, 82–96. doi: 10.1080/09581596.2021.1885628

Lee, D. H., and Jung, T.-P. (2020). A virtual reality game as a tool to assess physiological correlations of stress. *arXiv preprint* arXiv:2009.14421. doi: 10.48550/arXiv.2009.14421

Lee, M.-C. (2020). Free the data from the birdcage: Opening up data and crowdsourcing activism in taiwan. *PoLAR* 43, 247–264. doi: 10.1111/plar.12371

London, C. M. (1998). *Dunant's Dream: War, Switzerland and the History of the Red Cross*. Taylor and Francis.

Longstaff, P. H., and Yang, S.-U. (2008). Communication management and trust: their role in building resilience to "surprises" such as natural disasters, pandemic flu, and terrorism. *Ecol. Soc*. 13, 130103. doi: 10.5751/ES-02232-130103

Lykourentzou, I., Antoniou, A., Naudet, Y., and Dow, S. P. (2016). "Personality matters: balancing for personality types leads to better outcomes for crowd teams," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (New York, NY: Association for Computing Machinery), 260–273.

Lykourentzou, I., Vergados, D. J., Papadaki, K., and Naudet, Y. (2013). "Guided crowdsourcing for collective work coordination in corporate environments," in *International Conference on Computational Collective Intelligence* (Springer), 90–99.

Ma, J., Peng, Y., and Wu, B. (2021). Challenging or hindering? the roles of goal orientation and cognitive appraisal in stressor-performance relationships. *J. Organ. Behav*. 42, 388–406. doi: 10.1002/job.2503

Majchrzak, A., and More, P. H. (2011). Emergency! web 2.0 to the rescue! *Commun. ACM* 54, 125–132. doi: 10.1145/1924421.1924449

Mammadov, S. (2022). Big five personality traits and academic performance: a meta-analysis. *J. Pers*. 90, 222–255. doi: 10.1111/jopy.12663

Marc Cieslak, T. G. (2022). *Ukraine: How Crowdsourcing Is Rescuing People From the War Zone*. Available online at: https://www.bbc.com/news/technology-60785339. (accessed May 01, 2022).

Mark, G., Kun, A. L., Rintel, S., and Sellen, A. (2022). Introduction to this special issue: the future of remote work: responses to the pandemic. *Hum. Comput. Interact*. 1–7. doi: 10.1080/07370024.2022.2038170. [Epub ahead of print].

Martella, C., Li, J., Conrado, C., and Vermeeren, A. (2017). On current crowd management practices and the need for increased situation awareness, prediction, and intervention. *Saf. Sci*. 91, 381–393. doi: 10.1016/j.ssci.2016.09.006

Martins, L. L., and Shalley, C. E. (2011). Creativity in virtual work: effects of demographic differences. *Small Group Res*. 42, 536–561. doi: 10.1177/1046496410397382

Masthoff, J. (2004). "Group modeling: selecting a sequence of television items to suit a group of viewers," in *Personalized Digital Television* (Dordreht: Springer), 93–141.

Matsumoto, D., and Juang, L. (2016). *Culture and Psychology*. Boston, MA: Cengage Learning.

McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *J. Pers. Soc. Psychol*. 52, 1258. doi: 10.1037/0022-3514.52.6.1258

Mckinney Jr, E. H., Barker, J. R., Davis, K. J., and Smith, D. (2005). How swift starting action teams get off the ground: what united flight 232 and airline flight crews can tell us about team communication. *Manag. Commun. Q*. 19, 198–237. doi: 10.1177/0893318905278539

McManus, I., Keeling, A., and Paice, E. (2004). Stress, burnout and doctors' attitudes to work are determined by personality and learning style: a twelve year longitudinal study of uk medical graduates. *BMC Med*. 2, 29. doi: 10.1186/1741-7015-2-29

McMichael, M., Beverly, M., Noon, J., Patterson, T., and Webb, G. R. (1999). *Role Improvising Under Conditions of Uncertainty: A Classification of Types*. Disaster Research Center.

Mendonça, D. (2007). Decision support for improvisation in response to extreme events: learning from the response to the 2001 world trade center attack. *Decis Support. Syst*. 43, 952–967. doi: 10.1016/j.dss.2005.05.025

Mitchell, S. S., and Lim, M. (2018). Too crowded for crowdsourced journalism: Reddit, portability, and citizen participation in the syrian crisis. *Can. J. Commun*. 43, a3377. doi: 10.22230/cjc.2019v44n3a3377

Mohammed, S., and Angell, L. C. (2003). Personality heterogeneity in teams: Which differences make a difference for team performance? *Small Group Res*. 34, 651–677. doi: 10.1177/1046496403257228

Muethel, M., Gehrlein, S., and Hoegl, M. (2012). Socio-demographic factors and shared leadership behaviors in dispersed teams: implications for human resource management. *Hum. Resour. Manag*. 51, 525–548. doi: 10.1002/hrm.21488

Muhren, W. J., van de Walle, B. A., Muhren, W. J., and Van de Walle, B. (2010). Sense-making and information management in emergency response. *Bull. Am. Soc. Inf. Sci. Technol*. 36, 30–33. doi: 10.1002/bult.2010.1720360509

Myers, I. B. (1962). *The Myers-Briggs Type Indicator: Manual.* Consulting Psychologists Press

National Research Council. (2015). *Enhancing the Effectiveness of Team Science*. National Academic Press.

Neuman, G. A., Wagner, S. H., and Christiansen, N. D. (1999). The relationship between work-team personality composition and the job performance of teams. *Group Organ. Manag*. 24, 28–45. doi: 10.1177/1059601199241003

Normark, M. (2002). "Sense-making of an emergency call: possibilities and constraints of a computerized case file," in *Proceedings of the Second Nordic Conference on Human-Computer Interaction*. (Aarhus), 81–90.

Odo, C., Masthoff, J., and Beacham, N. (2019a). "Group formation for collaborative learning," in *International Conference on Artificial Intelligence in Education*. (Chicago, IL: Springer), 206–212.

Odo, C., Masthoff, J., and Beacham, N. A. (2019b). "Adapting online group formation to learners' conscientiousness, agreeableness and ability," in *SLLL@ AIED*. (Chicago, IL), 1–7.

Okolloh, O. (2009). Ushahidi, or 'testimony: web 2.0 tools for crowdsourcing crisis information. *Participat. Learn. Action* 59, 65–70.

O'Leary, M. B., and Mortensen, M. (2010). Go (con) figure: Subgroups, imbalance, and isolates in geographically dispersed teams. *Organ. Sci*. 21, 115–131. doi: 10.1287/orsc.1090.0434

Olson, J. S., and Kellogg, W. A. (2014). *Ways of Knowing in HCI, Vol. 2*. London: Springer.

O'Neill, T. A., and Allen, N. J. (2011). Personality and the prediction of team performance. *Eur. J. Pers*. 25, 31–42. doi: 10.1002/per.769

Pajonk, F.-G., Andresen, B., Schneider-Axmann, T., Teichmann, A., Gärtner, U., Lubda, J., et al. (2011). Personality traits of emergency physicians and paramedics. *Emerg. Med. J*. 28, 141–146. doi: 10.1136/emj.2009.083311

Palen, L., Hiltz, S. R., and Liu, S. B. (2007). Online forums supporting grassroots participation in emergency preparedness and response. *Commun. ACM* 50, 54–58. doi: 10.1145/1226736.1226766

Perez, A. J., and Zeadally, S. (2019). A communication architecture for crowd management in emergency and disruptive scenarios. *IEEE Commun. Mag*. 57, 54–60. doi: 10.1109/MCOM.2019.1800626

Pettersson, M., Randall, D., and Helgeson, B. (2004). Ambiguities, awareness and economy: a study of emergency service work. *Comput. Support. Cooperative Work* 13, 125–154. doi: 10.1023/B:COSU.0000045707.37815.d1

Pettet, G., Baxter, H., Vazirizade, S. M., Purohit, H., Ma, M., Mukhopadhyay, A., et al. (2022). Designing decision support systems for emergency response: Challenges and opportunities. *arXiv preprint* arXiv:2202.11268. doi: 10.48550/arXiv.2202.11268

Pfaff, M. S. (2012). Negative affect reduces team awareness: the effects of mood and stress on computer-mediated team communication. *Hum. Factors* 54, 560–571. doi: 10.1177/0018720811432307

Pittenger, D. J. (2005). Cautionary comments regarding the myers-briggs type indicator. *Consult. Psychol. J*. 57, 210. doi: 10.1037/1065-9293.57.3.210

Poblet, M., García-Cuesta, E., and Casanovas, P. (2013). "Crowdsourcing tools for disaster management: a review of platforms and methods," in *International Workshop on AI Approaches to the Complexity of Legal Systems.* (Heidelberg: Springer), 261–274.

Potts, L. (2013). *Social Media in Disaster Response: How Experience Architects can Build for Participation*. New York, NY: Routledge.

Quarantelli, E. L. (1988). Disaster crisis management: a summary of research findings. *J. Manag. Stud.* 25, 373–385. doi: 10.1111/j.1467-6486.1988.tb00043.x

Rammstedt, B., and John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the big five inventory in english and german. *J. Res. Pers*. 41, 203–212. doi: 10.1016/j.jrp.2006.02.001

Reuter, C., Ludwig, T., and Pipek, V. (2014). *Ad-hoc* participation in situation assessment: supporting mobile collaboration in emergencies. *ACM Trans. Comput. Hum. Interact*. 21, 1–26. doi: 10.1145/2651365

Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., and Laredo, J. A. (2013). Crisistracker: crowdsourced social media curation for disaster awareness. *IBM J. Res. Dev*. 57, 4–1. doi: 10.1147/JRD.2013.2260692

Rolland, J.-P. (2002). "The cross-cultural generalizability of the five-factor model of personality," in *The Five-Factor Model of Personality Across Cultures. International and Cultural Psychology Series*, eds R. R. McCrae and J. Allik (Boston, MA: Springer).

Ruef, M., Aldrich, H. E., and Carter, N. M. (2003). The structure of founding teams: homophily, strong ties, and isolation among us entrepreneurs. *Am. Sociol. Rev*. 68, 195–222. doi: 10.2307/1519766

Sabo, R., and Rajčáni, J. (2017). Designing the database of speech under stress. *J. Linguist. Jazykovednỳ Casopis* 68, 326–335. doi: 10.1515/jazcas-2017-0042

Salas, E., Driskell, J. E., and Hughes, S. (1996). *The Study of Stress and Human Performance. Stress and human performance(A 97-27090 06-53)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Salas, E., Tannenbaum, S. I., Kozlowski, S. W., Miller, C. A., Mathieu, J. E., and Vessey, W. B. (2015). Teams in space exploration: a new frontier for the science of team effectiveness. *Curr. Dir. Psychol. Sci*. 24, 200–207. doi: 10.1177/0963721414566448

Salehi, N., McCabe, A., Valentine, M., and Bernstein, M. (2017). "Huddler: Convening stable and familiar crowd teams despite unpredictable availability," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1700–1713.

Schilpzand, M., Herold, D., and Shalley, C. (2010). Members' openness to experience and teams' creative performance. *Small Group Res*. 41, 56–76. doi: 10.1177/1046496410377509

Schmidt, A., Wolbers, J., Ferguson, J., and Boersma, K. (2018). Are you ready2help? conceptualizing the management of online and onsite volunteer convergence. *J. Conting. Crisis Manag*. 26, 338–349. doi: 10.1111/1468-5973.12200

Schneider, R. O. (2011). Climate change: an emergency management perspective. *Disaster Prevent. Manag*. 20, 53–62. doi: 10.1108/09653561111111081

Sen, A. (1986). Social choice theory. *Handbook Math. Econ*. 3, 1073–1181. doi: 10.1016/S1573-4382(86)03004-7

Senot, C., Kostadinov, D., Bouzid, M., Picault, J., Aghasaryan, A., and Bernier, C. (2010). "Analysis of strategies for building group profiles," in *International Conference on User Modeling, Adaptation, and Personalization* (Heidelberg: Springer), 40–51.

Sharma, R. R., Chawla, S., and Karam, C. M. (2021). "Global gender gap index: world economic forum perspective," in *Handbook on Diversity and Inclusion Indices* (Edward Elgar Publishing).

Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations*. New York, NY: Penguin.

Smirnov, A., Levashova, T., and Shilov, N. (2011). "Ubiquitous computing in emergency: role-based situation response based on self-organizing resource network," in *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (Miami Beach, FL: IEEE), 94–101.

Smith, K. A., Dennis, M., Masthoff, J., and Tintarev, N. (2019). A methodology for creating and validating psychological stories for conveying and measuring psychological traits. *User Model Useradapt Interact*. 29, 573–618. doi: 10.1007/s11257-019-09219-6

Song, Z., Zhang, H., and Dolan, C. (2020). Promoting disaster resilience: operation mechanisms and self-organizing processes of crowdsourcing. *Sustainability* 12, 1862. doi: 10.3390/su12051862

Stachowski, A. A., Kaplan, S. A., and Waller, M. J. (2009). The benefits of flexible team interaction during crises. *J. Appl. Psychol*. 94, 1536. doi: 10.1037/a0016903

Stehr, N. (2001). *The Fragility of Modern Societies: Knowledge and Risk in the Information Age*. London: Sage.

Stein, R., and Swan, A. B. (2019). Evaluating the validity of myers-briggs type indicator theory: a teaching tool and window into intuitive psychology. *Soc. Pers. Psychol. Compass* 13, e12434. doi: 10.1111/spc3.12434

Stephens, K. K., and Robertson, B. W. (2022). 12 social media platforms and broader participation in crisis communication. *Soc. Media Crisis Commun*. 2nd edition. 12. doi: 10.4324/9781003043409-18

Stuster, J. W. (2011). *Bold Endeavors: Lessons From Polar and Space Exploration*. Annapolis, MD: Naval Institute Press.

Su, B.-C., Widjaja, A. E., and Chen, J. V. (2012). "Stress in virtual team vs face-to-face team: Is working in virtual team more stressful than face-to-face team?," *Conference paper). Consultion.*

Subramaniam, C., Ali, H., and Shamsudin, F. M. (2010). Understanding the antecedents of emergency response: a proposed framework. *Disaster Prevent. Manag*. 19, 571–581. doi: 10.1108/09653561011091904

Tarafdar, M., and Stich, J.-F. (2021). "Virtual work, technology and wellbeing," in *The SAGE Handbook of Organizational Wellbeing* (London: SAGE), 159–169.

Thompson, E. R. (2008). Development and validation of an international english big-five mini-markers. *Pers. Individ. Dif.* 45, 542–548. doi: 10.1016/j.paid.2008.06.013

Thoms, P., Moore, K. S., and Scott, K. S. (1996). The relationship between self-efficacy for participating in self-managed work groups and the big five personality dimensions. *J. Organ. Behav*. 17, 349–362. doi: 10.1002/(SICI)1099-1379(199607)17:4andlt;349::AID-JOB756andgt;3.0.CO;2-3

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston, MA: Houghton Mifflin.

Tuckman, B. W., and Jensen, M. A. C. (1977). Stages of small-group development revisited. *Group Organ. Stud*. 2, 419–427. doi: 10.1177/105960117700200404

Van de Ven, A. H., Delbecq, A. L., and Koenig Jr, R. (1976). Determinants of coordination modes within organizations. *Am. Sociol. Rev*. 41, 322–338. doi: 10.2307/2094477

van de Water, H., Ahaus, K., and Rozier, R. (2008). Team roles, team balance and performance. *J. Manag. Dev*. 27, 499–512. doi: 10.1108/02621710810871817

Velev, D., and Zlateva, P. (2012). Use of social media in natural disaster management. *Int. Proc. Econ. Dev. Res*. 39, 41–45.

Vivacqua, A. S., and Borges, M. R. (2012). Taking advantage of collective knowledge in emergency response systems. *J. Netw. Comput. Appl*. 35, 189–198. doi: 10.1016/j.jnca.2011.03.002

Wagner, S. L., Martin, C. A., and McFee, J. A. (2009). Investigating the "rescue personality". *Traumatology* 15, 5–12. doi: 10.1177/1534765609338499

Waller, M. J., Gupta, N., and Giambatista, R. C. (2004). Effects of adaptive behaviors and shared mental models on control crew performance. *Manag. Sci*. 50, 1534–1544. doi: 10.1287/mnsc.1040.0210

Wauben, L., Dekker-van Doorn, C., Van Wijngaarden, J., Goossens, R., Huijsman, R., Klein, J., et al. (2011). Discrepant perceptions of communication, teamwork and situation awareness among surgical team members. *Int. J. Quality Health Care* 23, 159–166. doi: 10.1093/intqhc/mzq079

Weick, K. E. (1993). The collapse of sensemaking in organizations: the mann gulch disaster. *Administrat. Sci. Q*. 38, 628–652. doi: 10.2307/2393339

Weisberg, Y. J., DeYoung, C. G., and Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the big five. *Front. Psychol*. 2, 178. doi: 10.3389/fpsyg.2011.00178

Wildman, J. L., Shuffler, M. L., Lazzara, E. H., Fiore, S. M., Burke, C. S., Salas, E., et al. (2012). Trust development in swift starting action teams: a multilevel framework. *Group Organ. Manag*. 37, 137–170. doi: 10.1177/1059601111434202

Worchel, S., and Shackelford, S. L. (1991). Groups under stress: the influence of group structure and environment on process and performance. *Pers. Soc. Psychol. Bull*. 17, 640–647. doi: 10.1177/0146167291176006

Yang, L., Prasanna, R., and King, M. (2009). On-site information systems design for emergency first responders. *J. Inf. Technol. Theory Appl*. 10, 2.

Yeo, J., Knox, C. C., and Jung, K. (2018). Unveiling cultures in emergency response communication networks on social media: following the 2016 louisiana floods. *Quality Quantity* 52, 519–535. doi: 10.1007/s11135-017-0595-3

Yu, T., Sengul, M., and Lester, R. H. (2008). Misery loves company: the spread of negative impacts resulting from an organizational crisis. *Acad. Manag. Rev*. 33, 452–472. doi: 10.5465/amr.2008.31193499

Yuan, F., and Liu, R. (2018). Feasibility study of using crowdsourcing to identify critical affected areas for rapid damage assessment: hurricane matthew case study. *Int. J. Disaster Risk Reduct*. 28, 758–767. doi: 10.1016/j.ijdrr.2018.02.003

Zhang, H. (2022). A real-time optimisation for disaster-relief distribution inheterogeneous crowdsourcing. *Impact* 2022, 23–25. doi: 10.21820/23987073.2022.1.23

Zhang, Y.-L., and Lu, C.-Q. (2009). Challenge stressor-hindrance stressor and employees work–related attitudes, and behaviors: the moderating effects of general self-efficacy. *Acta Psychol. Sin*. 41, 501. doi: 10.3724/SP.J.1041.2009.00501

Zijlstra, F. R., Waller, M. J., and Phillips, S. I. (2012). Setting the tone: early interaction patterns in swift-starting teams as a predictor of effectiveness. *Eur. J. Work Organ. Psychol*. 21, 749–777. doi: 10.1080/1359432X.2012.690399

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/frai.2022.818562

Crowdsourcing Team Formation With Worker-Centered Modeling

Federica Lucia Vinella, Jiayuan Hu, Ioanna Lykourentzou and Judith Masthoff*

Human Centred-Computing, Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands

Modern crowdsourcing offers the potential to produce solutions for increasingly complex tasks requiring teamwork and collective labor. However, the vast scale of the crowd makes forming project teams an intractable problem to coordinate manually. To date, most crowdsourcing collaborative platforms rely on algorithms to automate team formation based on worker profiling data and task objectives. As a top-down strategy, algorithmic crowd team formation tends to alienate workers causing poor collaboration, interpersonal clashes, and dissatisfaction. In this paper, we investigate different ways that crowd teams can be formed through three team formation models namely bottom-up, top-down, and hybrid. By simulating an open collaboration scenario such as a hackathon, we observe that the bottom-up model forms the most competitive teams with the highest teamwork quality. Furthermore, we note that bottom-up approaches are particularly suitable for populations with high-risk appetites (most workers being lenient toward exploring new team configurations) and high degrees of homophily (most workers preferring to work with similar teammates). Our study highlights the importance of integrating worker agency in algorithm-mediated team formation systems, especially in collaborative/competitive settings, and bears practical implications for large-scale crowdsourcing platforms.

Keywords: crowdsourcing, agent based modeling, social computing, self-organization, team formation

OPEN ACCESS

Edited by:
Alessandro Bozzon,
Delft University of Technology,
Netherlands

Reviewed by:
Matthias Hirth,
Technische Universität Ilmenau,
Germany
Andrea Mauri,
Delft University of Technology,
Netherlands

****Correspondence:***
Federica Lucia Vinella
f.l.vinella@uu.nl

Specialty section:
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

Received: *19 November 2021*
Accepted: *25 April 2022*
Published: *27 May 2022*

Citation:
Vinella FL, Hu J, Lykourentzou I and Masthoff J (2022) Crowdsourcing Team Formation With Worker-Centered Modeling. Front. Artif. Intell. 5:818562. doi: 10.3389/frai.2022.818562

1. INTRODUCTION

Online, on-demand, and large-scale work, also called crowd work, is increasingly gaining traction. For more and more people, this new labor model is no longer used just for side "gigs" but as a primary source of income. Companies are also shifting toward elastic labor models, increasing their share of crowd workers in favor of a full-time workforce (LLP, 2020). The pandemic accelerated this trend, forcing many people to re-skill, up-skill, and to work with unfamiliar and distant collaborators, especially in the form of crowd work (Barnes et al., 2015; De Stefano, 2015; Manyika et al., 2016). Besides small, straightforward tasks, also known as micro-tasks (Difallah et al., 2015), such as image recognition, captcha annotation, and translation, crowds are now increasingly being involved in generating solutions to difficult or "wicked" problems, such as climate change mitigation, disease spread prevention, or rapid innovation generation. Tasks of this sort also called macro-tasks (Khan et al., 2019), tend to be complex and ill-structured, with multiple knowledge interdependencies and no straightforward solution. Because of their complex and open-ended character, these tasks typically require collaboration among workers of different skill sets and knowledge backgrounds. While micro-tasks lend themselves to being solved quickly and are therefore short-lived and affordable, macro-tasks frequently urge interdisciplinary collaboration, require more time, and are more challenging due to their breadth of scope.

Driven by the need to innovate and stay ahead of competition, companies increasingly make open calls for solving creative challenges through platforms such as OpenIdeo (Lakhani et al., 2012) and InnoCentive (Lakhani and Lonstein, 2008), where teams of crowd workers compete for prizes (Betts and Bloom, 2014). Another type of commercial task, which highly depends on the successful collaboration of crowd teams are online creative hackathons, for example those dedicated to video game development. Events such as the Global Game Jam gather thousands of online participants, including artists, developers, marketers, who form teams to compete for the best game product; sustainable game production in this case directly depends on the participants' ability to find the right group to work with Whitson et al. (2021). Aside from pure commercial interest, crowd team formation is also at the core of governmental initiatives. With the profound societal changes brought by the COVID-19 pandemic, grassroots entrepreneurship efforts have increased to stimulate economies and slow down infection rates. With 9,000 participants from 142 countries and 49 states, the Massachusetts Institute of Technology (MIT) COVID-19 Challenge is the most recent exemplary attempt addressing immediate needs with rapid innovation through a series of virtual hackathons involving *ad-hoc* teams of remote participants (Ramadi and Nguyen, 2021).

To coordinate the efforts of such workforce, crowdsourcing research has started to look into team formation algorithms as automated, scalable solutions. Routinely, team formation algorithms match workers according to objectives such as interpersonal compatibility (Lykourentzou et al., 2021) and social network connectivity (Liu et al., 2015; Rahman et al., 2019). One of the limits of computed team formation solutions—which we address in this study—is the omission of the workers' preferences and evolving relationships in the algorithmic objective function. In other words, workers have no say in whether they want to stay in a team chosen for them, and who they will work with. Team formation algorithms usually collect the workers' profile features *before* the task begins (Liu et al., 2015; Rahman et al., 2019), but then do not adjust to the workers' utilities and pay-offs *during* the collaboration. Although the workers' attributes are gathered only once, they are often assumed to suffice for the formulation of optimal teams. As a result, algorithms often fail to capture covert features such as temporal team dynamics information, collaboration preferences, intra-group compatibility, and individual risk appetites; features that play a key role in teamwork success (Degli Antoni et al., 2021). Aside from profiling information, team formation systems have recently started to factor social network properties in their objective functions, bringing together teams based on their network tie strength (Salehi and Bernstein, 2018) mutating as the collaboration evolves. However, in this case too, the system does not adapt its decisions based on worker feedback concerning the enforced rotations, and it does not account for cases where the workers' ties deteriorate or even break. In reality, however, individual team member agency makes up a significant portion of whether a team will be able to perform successfully or not, and removing it could mean reducing the adequacy and fairness of the team formation system.

Concerns about the poor representation of worker agency in automated team formation solutions are starting to surface. Recent research shows that purely top-down solutions result in rigid team structures and workflows that stifle creativity and initiative-taking, and inhibit workers from adapting their problem-solving strategy to the task needs, which, in turn, is detrimental for complex and open-ended tasks (Retelny et al., 2017). Forcing workers to work with specific people can also cause psychological fatigue and discomfort, reduce user autonomy, alienate workers, and lead to less-than-optimal collaboration (Rasmussen and Jeppesen, 2006; Lawler and Worley, 2009). A growing number of studies are starting to propose ways to incorporate worker agency, including preferences but also unconscious drivers, into crowd work settings, so as to directly and positively advance teamwork quality, efficiency, and well-being. Gaikwad et al. (2015, 2017) and Whiting et al. (2017) show that incorporating elements of open governance has been found to promote trust between workers and task providers. Yin et al. (2018) show that trusting workers with the work schedule increases the number of tasks completed without compromising quality, with workers actually willing to forego significant pay to control their working time. Specifically to the domain of collaborative work, Lykourentzou et al. (2016b) use a technique known as team dating, where people meet with candidate teammates in rapid succession before deciding to settle into teams. Although their solution integrates agency only indirectly, by forming teams based on peer evaluations of the intermediate team dates, this study shows that accounting for worker feedback during team formation can have a positive effect on team performance and satisfaction. Looking at research preceding the online crowdsourcing and open collaboration movements (Jackson, 1983; De Dreu and West, 2001), we also spot fundamental evidence on the importance of allowing workers' agency in teamwork such as through minority dissent and participation in decision making. Granted autonomy, individuals not only produce improved results (Gilson and Shalley, 2004; Costa et al., 2018), but also exhibit healthier mental states associated with self-governance, feelings of empowerment, reduced stress, sense of ownership over their work and ideas, and increased group interdependence and cohesion (Carless and De Paola, 2000; Rasmussen and Jeppesen, 2006; Haas and Mortensen, 2016). In this study we are interested in exploring how more worker-centered and bottom-up team formation compares to the prevalent approach of forming teams in a top-down and purely algorithm-driven manner. We do so by modeling and comparing three team formation systems, namely a (i) **fully bottom-up** system, where we model algorithm involvement to be minimal and team formation to lie almost exclusively on worker decisions, a (ii) **fully top-down** one, where we adapt a latest state-of-the-art team formation algorithm (Salehi and Bernstein, 2018), (iii) and (iv) a **hybrid** system, which borrows elements from the previous two. Although the three system models all aim to tackle team formation, their difference lies in the level of agency they permit and the degree of algorithmic mediation they enforce during team formation.

The fully bottom-up system (which we call SOT from Self-Organized Teams) is represented by two models. The first model, called **Radical SOT (R-SOT)**, prioritizes individual worker over team preferences of new teammates, and dismantles an existing team if at least one of its members decides to leave. The model focuses on facilitating novel interactions between the workers and leads to radical restructures of the collaboration network. The second model, called **Conservative SOT (C-SOT)**, facilitates bottom-up team formation in a less radical manner, since it prioritizes team over solo worker agency. In this model, teams looking for members have priority over individuals, and a team remains together as long as two of its members wish to keep collaborating. This model prioritizes majority consent over minority dissent. For the top-down model, we adopt **Hive**, a community-based team formation algorithm by Salehi and Bernstein (2018). Hive was chosen as it is a state-of-the-art algorithm and it represents the latest trend in top-down team formation approaches which adapt their decisions during the task rather than making them only once in the beginning. Briefly, Hive uses social network information to rotate people across teams so as to balance tie strength and network efficiency, and computes teamwork quality whilst rotating teams according to a stochastic search suited to minimize algorithm complexity. Finally, combining bottom-up and top-down approaches, we propose and add to the comparison a third hybrid system model named **HiveHybrid**. The model combines worker agency with algorithmic mediation. In this model, the algorithm offers to rotate workers according to the Hive system's objective function, but workers have then the option to accept or decline these proposals based on whether they are predisposed to break ties with their teammates or not (depending on their assessment of team reward and their personal risk appetite). In HiveHybrid, the workers' preferences play as much of a role in team formation as the coordinating algorithm. We run a comparative study, using agent-based simulations on the scenario of team formation for a creative game development hackathon, to evaluate differences in teamwork quality across these three team formation models. We focus on answering the following research questions:

1. **RQ1. How does bottom-up team formation compare with top-down and hybrid approaches?** We first compare the three team formation system models on the teamwork quality they yield, since quality is the primary and typical concern of crowdsourcing research, platforms, and clients. We use three metrics, namely the best, average, and worst teamwork quality, which are relevant depending on the requirements and constraints of the specific crowd work use case one is interested in.
2. **RQ2: How do population behavioral tendencies affect the outcome of bottom-up online teamwork?** Since bottom-up systems are more influenced by the participating workers' attributes, tendencies, preferences, and decisions than top-down ones, we systematically evaluate the effects of certain worker population's attributes on team performance. The objective of this evaluation is to help future crowdsourcing systems design incentives or countermeasures for different expected population behaviors, concerning team exploration tendencies, population size, and tendencies toward teamwork diversity. To systematically evaluate the effects of each of these attributes, we break down this research question into the following three sub-questions:
   - **RQ2.1: How do different risk appetites affect teamwork output in bottom-up models?** Workers with a high risk appetite tend to leave teams and rotate more often (preference for exploration) compared to workers with a lower risk appetite who tend to form more lasting teams (preference for exploitation). Risk appetite is expected to affect teamwork quality as it affects the number and structure of the self-organized teams. As a personal attribute, risk appetite is not only to influence the frequency of changes but also the preference of tasks (i.e., some workers might prefer tasks that are higher paid but less likely to be completed successfully), however, for simplicity, we have focused on one task type for this study.
   - **RQ2.2: How do different worker population sizes affect teamwork output in bottom-up models?** Evaluating the effects of changes in the population size helps to understand how changes in crowdsourcing collaborative participation affects the workers' search space, coordination costs, and teamwork quality.
   - **RQ2.3: How does homophily, i.e., the tendency to prefer working with similar teammates, affect teamwork output in bottom-up models?** Homophily is known to affect social interactions as people tend to choose (work with) partners based on shared physical and cultural cues (Haun and Over, 2015). Evaluating the effects of different homophily thresholds of the participating worker population on quality can facilitate the evaluation of whether certain explicit system incentives are needed to encourage workers to join forces with different collaborators or not.

Our results contribute to the development of future crowdsourcing tools for team formation that can be adapted—with the introduction of more or less degrees of agency—to the needs of the particular use case and the characteristics of the specific worker population involved. For one, we observe that **self-organization supports the formation of competitive teams**. In use case scenarios where innovation is key, a system capable of preserving worker agency can be a good return-on-investment for organizations that leverage competitive skills. Inspiring exploration across a large pool of curious workers seems to be an adequate strategy for forming competitive teams in bottom-up settings; so is the emancipation of team similarity where workers favor teammates of similar cultural and demographic attributes when workers have full control over team rotation. On the contrary, usage scenarios where it is more important to maintain fairness than performance, could benefit more from algorithmic-mediated team formation solutions to explicitly moderate the segregating tendencies we observe in fully bottom-up models. In this case, our results indicate that a hybrid system such as HybridHive constitutes an advantage over either fully top-down or fully bottom-up models, since it balances the global distribution of resources with worker agency mediating micro behavior through macro structures.

The rest of this paper is organized as follows. We first provide an overview of existing team formation approaches focusing on the collaborative crowdsourcing domain (Section 2). Afterward, we dive deeper into the modeling components that make the three team formation systems examined in this study (Section 3). Next, we present the results of the simulations comparing the three systems, mapped to the relevant research questions (Section 4). We then proceed by discussing the applicability and relevance of the findings (Section 5), followed by reasoning on the limitations of this study (Section 6). We conclude the paper with the main findings, key messages, and final remarks (Section 7).

2. RELATED WORK

2.1. Team Formation Algorithms for Managing Online Work

Broadly speaking, the Team Formation Problem (TFP) is the problem of allocating a set of people to subsets, referred to as teams, according to a set of criteria that vary depending on the application area (Juárez et al., 2021). As illustrated in Juárez et al. (2021)'s recent review and taxonomy, TFP research has been persistently increasing over the past ten years. The problem encompasses a wide variety of applications, ranging from the assignment of students to study groups, to the distribution of patients to hospital rooms, and from the assignment of reviewers to papers, to the composition of teams for collaborative work purposes. In this paper, we focus on team formation for online work and, in particular, large-scale crowd participation in collaborative work. The research community has mostly focused on designing algorithms that ensure the quality of digital work by orchestrating people in a *top-down* manner, mainly with the objective to optimize costs. A recent extensive bibliometric analysis of 268 articles on crowd work task recommendation (Yin et al., 2020), covering the period of 2006–2019 (practically since the onset of crowd work) confirms the above, revealing that the largest and most durable research clusters focus on forming teams to optimize the task's budget, using methods such as dynamic programming, routing, and allocation. Similar methods are standard practice in operational research (Taha, 2013), an area traditionally geared toward optimizing supply chain management and manufacturing.

2.1.1. Static Team Formation Models: Making Decisions Only Once

The problem of forming optimal teams is generally $\mathcal{NP}$-hard, and for this reason the majority of team formation algorithms make their decisions in a deterministic fashion and only once **at the beginning of the task**. The algorithm's intervention in these cases ends with one-off team formation decisions, after which the teams remain stationary, indisputable, and irreversible. Commonly used team formation systems typically bank on pre-existing workers profiling data, such as skills, availability, or hourly wage to estimate teamwork dimensions including expertise complementary (Rahman et al., 2019), team costs (Liu et al., 2015), and team roles (Retelny et al., 2014; Valentine et al., 2017). Subsequently, the algorithms feed this data to machine learning or combinatorial optimization models to produce (near-)optimal solutions. An example of such an approach is the work by Rahman et al. (2019) proposing an algorithm that relies on worker skills, wage, and pairwise affinity to match workers with teams and teams with tasks. Other examples include the work by Yu et al. (2019) using the Hungarian algorithm to calculate matches based on skill, task complexity, and active time, and the work by Ahmed et al. (2020) exploring crowdsourcing sequential arrival with the objective to maximize teams' utility and diversity.

Besides handling team formation as a combinatorial optimization problem, there are other ways that crowdsourcing team formation problems have been thought of. An example is the work by Liu et al. (2015) operating through a mechanism design approach that proposes a task pricing algorithm seeking to assemble crowd teams on the basis of costs and skills. This work looks at worker truthfulness in the bidding process as a desirable property of the model, where incentive compatibility results in the preferred dominant strategy. Models of this kind rely on pre-calculated assumptions and deterministic predictions to make their team formation decisions and are especially useful in settings where task requirements are well-defined and known a priori, and worker characteristics are immutable. For these tasks, the use of pre-calculated teams permits to scale-up and compute solutions that are both computationally efficient and high-quality (Avis, 1983). However, static models do not appraise changes in the collaborative environment, for example, changes in the workers' preferences and affinities as they work together, the evolution of team dynamics, or changes in the task requirements (e.g., expertise needed) over the course of the collaboration (Ananny, 2016; Faraj et al., 2018). Consequentially, they risk creating rigid team structures that cannot optimally address tasks of evolving complexity.

2.1.2. Dynamic Team Formation Models: Adapting to Change

Recently, research has started looking into adaptive algorithms that make their team formation decisions **during the task**, as the collaboration unfolds. In this direction, Zhou et al. (2018) propose an algorithm using multi-armed bandits with temporal constraints, which explores the trade-offs among various dimensions of team structure, such as interaction patterns or hierarchies. By letting each bandit observe team performance and choose which arm to use next, the algorithm decides when and how to make changes in the structure of each team. In another example, Retelny et al. (2014) and Valentine et al. (2017) propose Foundry, a crowd management system that assembles workers into role-based teams. Although workers can request changes in the original teams, the final decision is made by a small number of experts and the task requester. Aside from skill sets, budget, and time, a small set of recent studies has started proposing team formation algorithms that harness social network qualities such as connectivity (Salehi and Bernstein, 2018), centrality (Hasteer et al., 2015), and marginality (Wang, 2020), as non-trivial parameters affecting teamwork performance across time. In this direction, Jiang et al. (2019) propose a team formation algorithm that instead of forming artificial teams, based on the individual teammates' skills, cost, or other features, utilizes groups that have been naturally organized through social networks, and allocated them to tasks in a priority-based manner based on their capacity to address the task. In the same line, Wu et al. (2021) propose a graph-based algorithm that estimates the accuracy of allocating a group of workers to a task, by joining the factorized matrixes of the workers' social network connections with their work history of on tasks.

Relevant to this study is the work of Salehi and Bernstein (2018). It envisages an online model (Hive algorithm) that balances two competing forces in team formation optimization: network efficiency and tie strength among the different worker pairs. It conceives crowdsourcing team formation as a graph partitioning problem where disjoint subsets (teams) benefit from strong ties but suffer from a lack of connectivity within the collaborative environment. This approach is an attempt to reconcile familiarity (obtained when relationships remain constant over time) and serendipity (spurred when breaking old ties and forming new ones). It handles team formation problems sequentially and in a stochastic fashion, juxtaposing top-down appointed team rotation with a series of collaborative stages of crowdsourcing work. It mediates team rotation by picking probabilistic moves at every round in keeping with a combination of tie strength and network efficiency. Rotating teams in crowd open collaboration resulted to be remarkably successful in connecting diverse perspectives. However, the same model provoked discomfort as workers could not determine by themselves the outcome of the match and could not depart from inefficient teams or decide to remain in the preferred one. We use Hive as a state-of-the-art representative benchmark of top-down work coordination in simulated scenarios. Although the above algorithms adapt to changes in team performance and task requirements that may occur over time, they are still fully top-down mechanisms that infer their decisions without actively engaging workers in the decision-making process.

In summary, relying on top-down coordination to form teams presents clear limitations. First, it limits the breadth of attainable work to tasks that the algorithm can decompose and assign to workers according to predefined criteria. For this reason, top-down crowdsourcing team formation solutions are ideal for tasks that are usually well-structured, with known interdependencies, and clear knowledge boundaries. However, for creative complex tasks and innovation generation they still tend to ignore worker self-organizing abilities and under-cater work flexibility. Subsequently, they fail at empowering crowd workers and drastically limit personal development opportunities (Roy et al., 2013; Schriner and Oerther, 2014). Ergo, another major limitation of top-down solutions—especially in crowdsourcing collaborative spaces—is the workers' confinement and isolation within the collaborative environment where algorithms direct and workers execute (Berg, 2015; Smith and Leberstein, 2015; Popescu et al., 2018; Gray and Suri, 2019). Furthermore, the pay-per-work model leads to the commodification of online work and online workers (Wood et al., 2019). It also means that workers must bear "work-for-labor" costs, i.e., costs for activities like breaks, training, or waiting for work—which are necessary to perform the task—but they are not part of the work itself (Berg, 2015; Florisson and Mandl, 2018) as they are still treated as separate entities from the collaboration and the end-result. For these reasons, ethical issues also arise (Silberman et al., 2018) concerning the labor conditions of crowd workers, their rights and legal status (Deitz, 2016), and "lock-in" phenomena where workers are tied to platform monopolies and non-transferrable profile information (e.g., performance history). In the last years, more and more researchers are raising critical voices (Smith and Leberstein, 2015; Gray and Suri, 2019) regarding the need to shift away from the canonical top-down crowdsourcing team formation systems and give workers agency, control, and self-determination capacity.

2.2. Self-Organization in Team Formation: Mediating Through Guidance

The term self-organization is present across several managerial and scientific fields spanning from software development communities to complex systems and natural science. The term describes the emergence of spontaneous processes and interactions between entities of originally disordered systems (Yates, 2012; Anzola et al., 2017). In team formation, self-organization usually describes the behavior of individuals as they form groups and collaborate autonomously and without pre-defined leadership. In software development, the term self-organization typically indicates the distribution of workload among teammates who flexibly shift responsibilities and partake in decision-making (Highsmith, 2009). Self-organized teams are known to benefit from transferable authority (Moe and Dingsøyr, 2008), as well as from robust and adaptable collaborative networks (Marzo Serugendo et al., 2003). The work of Lykourentzou et al. (2016b) explores the self-organization phenomena in the crowdsourcing domain in the way it affects teamwork. In their study, unfamiliar workers try out potential teammates before settling into teams, thus self-organizing into reciprocal work groups. Their results show that handing decision-making power to crowd workers increases performance compared to top-down team allocation. Further, as shown in Rokicki et al. (2015), when applying self-organization to crowd teams reward systems, ergo when allowing people to decide upon reward distribution, the self-governing approach results in fairer compensation than conventional top-down reward systems.

However, simply relying on self-organization as an emerging, non-controlled property is not enough for digital labor systems. For one, the need to adhere to financial and quality targets can suffer from purely self-organized means. Entirely autonomous teams can risk overspending on resources and coordination time, two essential aspects of teamwork. Consequently, we evaluate the efficacy of **guided self-organization** as a resolution between central control and self-governance. This relatively new approach (Prokopenko, 2009) aims to regulate self-organization in dynamic complex systems by combining task-independent global goals (e.g., autonomy, fairness, governance) with task-dependent constraints (e.g., costs, efficiency) on local interactions. Up to now, this approach has been thoroughly researched in robotics (Martius and Herrmann, 2012; Nurzaman et al., 2014). As for crowd work, guided self-organization is the golden mean between safeguarding worker autonomy

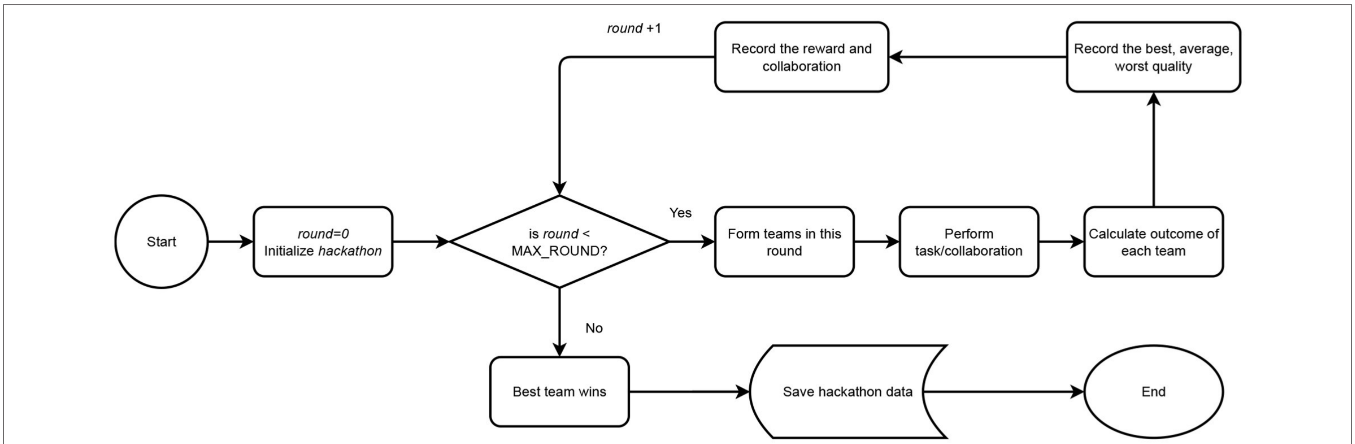

round +1
Record the reward and collaboration
Record the best, average, worst quality
Start
round=0
Initialize hackathon
is round < MAX_ROUND?
Yes
Form teams in this round
Perform task/collaboration
Calculate outcome of each team
No
Best team wins
Save hackathon data
End

FIGURE 1 | System architecture displaying the steps taken by the system in accordance with the hackathon design starting from the initialization of the agents and proceeding to the formation of teams assessed across ten rounds.

and protecting digital work platforms from disintermediation (Jarrahi et al., 2020). In the past, the principles of guided self-organization (albeit under a different name) have touched upon collaborative knowledge production (Lykourentzou et al., 2010) and crowdsourcing teams (Lykourentzou et al., 2019). These studies indicate that guided self-organization is a potentially effective coordination model for crowd collaboration in a manner that is distributed, efficient, and fair. In our study, guided self-organization is represented by a hybrid model which combines bottom-up self-organization with top-down community-based team formation.

3. METHODOLOGY

In this study, we attempt to re-create and predict emerging properties of online crowdsourcing collaborative settings where the actions of multiple workers—and the intervention of team formation approaches—affect teamwork and team output. Our simulation consists of three components: the **setting** (Section 3.1), the **agents** (Section 3.2), and the modeling of the **work coordination models** (Section 3.3). These are fundamental parts of the simulated scenario and exhibit behavioral properties, functional objectives, and constraints typically present in real-world crowd collaborative systems. **Figure 1** showcases the hackhathon system architecture.

3.1. Setting

Our simulation setting is a cycle-based online crowdsourcing hackathon. Online hackathons represent collaborative scenarios where several remote crowd workers of different backgrounds can gather in teams to create projects and compete for prizes. Even though hackathons have originated from the software development community (e.g., cybersecurity, game jams, open-source development, and operating systems) (Nolte et al., 2018), they are increasingly popular in other domains such as crowdsourcing innovation (Temiz, 2021; Wang et al., 2021). Further, as society faces progressively more global challenges, the help of citizens—and more broadly crowds—is also being used to find solutions to universal problems such as carbon emissions, household waste, and deforestation through collective idea generation (Monsef et al., 2021).

In our scenario, the (hypothetical) company recruits participants (game developers, marketers, designers, testers) online from popular crowdsourcing platforms (Amazon Mechanical Turk, Upwork, etc.) or other venues (e.g., creative hubs)[1] and retains them until the end of the event (Section 3.2). During the first round, workers are initially grouped randomly into teams of four and then they are required to collaborate for a number of consecutive rounds, which for our scenario is set to ten (Section 3.4). Depending on the approach involved in the team formation process (Section 3.3) and the level of the workers' agency modeled in the system, workers may move to other teams voluntarily or by top-down means. At the end of each round, each worker is given a reward (which can be thought of in monetary terms, e.g., in US Dollars), based on the ranking of their team's quality compared with other teams using the reward function (Equation 1).

$$reward = \frac{n-j}{n-1}, \text{ for team of ranking } j, \tag{1}$$

where n is the number of teams.

The product of each team is evaluated, using a quality function (described by Equation (2) and introduced in detail in Section 3.2.2, which simulates external evaluation by means of an external jury). At the end of the final (tenth in our simulations) round, the system automatically identifies the final best, average, and worst projects computed by means of the teamwork quality function (Equation 2).

[1] For this study, we chose Amazon Mechanical Turk as the example hiring platform since its demographics are well-known (Difallah et al., 2018). Nonetheless, we acknowledge that the population is expected to differ on platforms such as Upwork and other virtual creative hubs.

TABLE 1 | Worker attributes observing their mutability, visibility, type, possible value, and distribution.

Attribute	Possible attribute	Instantiation	Mutability	Visibility
Knowledge domain	Developer; Designer; Marketer; Tester	Random uniform	Immutable	Manifest
Nationality	USA; India; Other			
Educational level	High school; Bachelor; Master or above			
Age (years)	<20; 21–30; 31–40; 41–50; 51–60; >60	<20 = 2%, [21..30] = 40%, [31..40] = 36%, [41..50] = 7%, [52..60] = 9%, >60 = 4%		
Personality	Dominant; Inspiring; Supportive; Cautious	D = 50%; I = 10%; S = 20%; C = 20%		Latent
Risk appetite	[0,1]	Beta distribution ($\beta = 2$, $\alpha = 2$)		
Expertise	[0,1]	Beta distribution ($\beta = 2$, $\alpha = 2$)		

3.2. Agents

Here we describe the modeling of the two key strategic agents of the team formation problem, namely the: (i) workers and their individual characteristics and (ii) the teams, consisting of multiple workers.

3.2.1. Worker

For the simulation and in line with our working scenario, we focus on crowd worker profiles that can be involved in video game development in the context of a hackathon. We model worker attributes (**Table 1**) into two categories: (i) **manifest** (Section 3.2.1.1) and (ii) **latent** (Section 3.2.1.2) properties. Manifest attributes are those worker characteristics that are straightforwardly noticeable by others and can be captured into the profiling information of online team formation systems (Lykourentzou et al., 2021). These attributes are the workers' *knowledge domain, nationality, educational level*, and *age*. The latent attributes withhold worker characteristics that are not directly evident to others but that do affect the workers' compatibility, exploratory behavior, and competency. These latent characteristics are *personality, risk appetite*, and *expertise*. We distribute both manifest and latent attributes in relation to a set of probability functions based on previous work and modeled on the likelihood of occurring within a crowd population.

3.2.1.1. Manifest Attributes

1. **Knowledge domain**. This attribute captures worker expertise and is intended for the division of labor within a team. Following our working scenario on game development, we model four knowledge domains, namely: (1) Developer (typically a computer science specialist who creates software and application), (2) Designer (a game designer invested in software design, computer graphics, and animation), (3) Marketer (specialist in charge of monitoring market trends and creating advertising campaigns), and (4) Tester (worker in charge of playing the game to find errors and issues and evaluate the user experience). These domains are abstract representations of real-world work division in project-based teams and are relevant to scenarios where interdisciplinarity is vital to teamwork (Haeussler and Sauermann, 2020). All four knowledge domains manifest in the population with a random uniform distribution such that each trait has an equal probability of being expressed in the worker pool.
2. **Nationality**. This attribute imitates cultural differences in communication style, norms, and customs (Ortu et al., 2017) and may affect the workers' likelihood of seeking others similar to them (Centola et al., 2007). We model three nationalities as the most common among crowdsourcing workers (Difallah et al., 2018), namely: (1) USA, (2) Indian, and (3) Other nationalities. Just like the knowledge domain, nationalities are distributed randomly and uniformly across the population.
3. **Educational levels**. We model the workers' highest obtained educational qualification as: (1) High school, (2) Bachelor, or (3) Master or higher. We include the educational level in the working model for two main reasons. The first is that educational background is often a pivotal factor in hiring processes, including screening in crowdsourcing platforms such as AMT and Prolific (Prolific Team, 2021). The second reason is that, like social status, educational levels affect workers' preferences for teammates (McPherson et al., 2001) of similar or higher education. This attribute is also randomly and uniformly distributed in the worker pool.
4. **Age**. We model age in intervals [<20; 21–30; 31–40; 41–50; 51–60; >60] to classify differences in work culture, viewpoints, and collective identity. Age may also affect worker choice of teammates, with workers tending to favor collaborators of similar age with whom they are likely to share similar attitudes and beliefs (McPherson et al., 2001). The age attribute is distributed in accordance with crowdsourcing demographic statistics by Difallah et al. (2018) where <2% are younger than 20 years old, ∼40% are between 21 and 30, ∼36% are between 31 and 40, over 7% is between 41 and 50, a little over 9% is between 51 and 60, while the remaining 4% is older than 61.
5. **Past average reward**. We model past average reward as the average of the rewards received by the worker through their previous team collaborations (as a reminder a worker's past reward per round is calculated using Equation 1).

3.2.1.2. Latent Attributes

1. **Personality**. Using the DISC personality model by Marston (2013), we classify workers' approaches to leadership roles and team problems as being: (1) Dominant (D), (2) Inspiring (I),

(3) Supportive (S), or (4) Cautious (C). DISC was selected as it is widely used specifically in work-related settings, for example during hiring processes (Furlow, 2000). Each trait influences a worker's attitude to teamwork and mimics interpersonal factors affecting team processes. Based on the study by Lykourentzou et al. (2016a), we factor workers' personalities in the teamwork quality calculation and bonus teams of equally balanced personality traits (Equation 2). The aforementioned study also provides us with the distribution of personalities in a typical crowd work population, as follows: 50% of the workers are of personality type D, 20% are of type S, another 20% are of type C, and the remaining 10% are of type I.

2. **Risk appetite**. This represents to what extent workers are willing to explore new teams. The concept takes from the exploration-exploitation trade-off dilemma (Berger-Tal et al., 2014) concerning the problem of choosing between conserving a state or exploring new ones. In this case, a worker's risk appetite is mutable and determines one's tendency to seek collaborators outside their teams. We model each worker's risk appetite as value in the $[0, 1]$ range and distribute it across the population using a beta distribution probability function (Eugene et al., 2002). The beta distribution was chosen because it is bounded and can be easily modeled to illustrate various probability density functions (e.g., most workers having a low risk level with a long tail of high risk-workers, or vice versa).
3. **Expertise.** This attribute concerns the workers' level of ability in the knowledge domain in which they belong (Developer, Designer, Marketer, or Tester). It is not to be confused with education which is the formal training and schooling of the worker, which is used for computing the decisions taken by the workers on the basis of similarity. Expertise is modeled as a manifest attribute in the sense that, just like in real conditions, other workers (and the profiling system) can easily see *which* knowledge domain each other worker belongs to, but not *how good* the worker is in the specific domain. In our simulation, workers' expertise is treated as an immutable parameter and is distributed in the population with a beta probability distribution function (PDF) similar to a bell curve (with parameters $\alpha = 2$ and $\beta = 2$), i.e., most workers are of average expertise in their respective knowledge domains, and less workers are either complete novices or complete experts.
4. **Homophily**. This attribute describes the degree to which workers tend to prefer working with people that are more or less similar to themselves. We model homophily as it is one of the most studied motivators for forming social ties (McPherson et al., 2001). This principle structures human connections and knowledge exchange as well as restricting social worlds and interactions through subjective preferences for similar nationality, age, education, etc. (McPherson et al., 2001). We model worker's homphily as a cosine similarity score between two workers' vectors consisting of the attributes knowledge domain, nationality, educational level, and age.

3.2.2. Team

A team is a group of workers collaborating together for the duration of one or more rounds. Each team is a combination of the participating workers' attributes and their interactions, which affect the team output. Specifically, we model the output of each team, hereby referred to as **teamwork quality**, as a weighted sum of three elements, namely the team's: (1) skill, (2) interpersonal compatibility, and (3) size:

$$\begin{aligned}\textbf{Teamwork Quality} = {} & \pi \times \textbf{Team skill} + \mu \\ & \times \textbf{Interpersonal compatibility} \\ & + (1 - \pi - \mu) \times \textbf{Team size}, \end{aligned} \tag{2}$$

where:

1. **Team skill** is modeled as a weighted sum of the team members' expertise across the knowledge domains of the task, adjusted by a diminishing factor for repetitive expertise. Higher individual levels of expertise and higher coverage of the task's knowledge domains lead to higher team skill. We detail the modeling of the team skill element in Section 3.2.2.1.
2. **Interpersonal compatibility** is the degree to which the different teammates can work together harmoniously according to their work personality attribute. Higher coverage of the four personality types foreseen by the DISC test (D, I, S, and C) leads to higher teamwork quality. The presence of two or more members with personality type D (Dominant) lowers teamwork quality as it is known to produce clashes in collaborative crowd work settings (Lykourentzou et al., 2016a). We detail the modeling of this element in Section 3.2.2.2.
3. **Team size**. Team size affects teamwork quality, with teams above or below a certain threshold producing less-than-optimal results.

All three elements are measured in the $[0, 1]$ range, which also bounds teamwork quality in the same range. The coefficients π and μ can vary depending on the desired modeling. For our specific simulation, we set them to $\pi = 0.4$ and $\mu = 0.4$ (see Section 3.4).

3.2.2.1. Team Skill

Team skill is calculated as the combination of: (1) *coverage of the task's knowledge domains* by the members of the team, and (2) their *expertise* levels per domain. We assume that workers' expertise contributes positively to teamwork and that the workers' skill diversity promotes team interdisciplinarity. In case there are several teammates with the same knowledge domain in a team, we apply a diminishing factor to their skill utility in descending skill order. For example, in a team where three workers share the same domain, the second most expert in that domain has their skill utility discounted by a diminishing factor (which for our simulations is set to 0.10). All other lesser experienced workers of the same domain have their skill utility diminished by the same factor squared. We also discount 10% to all first-met teammates to account for the fact that the process of getting to know others and adjusting to new ways of working together taxes teamwork. Team skill is therefore calculated as follows:

$$\textbf{Team skill} = \frac{1}{s_t} \times \sum_{d=1}^{n} (\sum_{i=1}^{c_d} expertise_{d,i} \times \theta^{i-1}), \qquad (3)$$

where s_t is the size of the team, n is the number of total domains (four in this study), c_d is the number of workers in domain d, $\theta = 0.1$ is the diminishing factor for multiple expertise, and $expertise_{d,i}$ is the expertise of worker i in domain d.

3.2.2.2. Team Compatibility

We recognize the diversity of personality types as a representative measure of team interpersonal compatibility. More specifically, according to the DISC personality model, the more diverse and balanced a team is in regards to their DISC personalities, the more performant that team will be. To this end, the best team in our modeling is one the members of which cover all four DICS personality types. Such a team is optimal because it avoids both work disputes (which take place in the event of too many dominant types) and a lack of cohesion (which happens in case of missing personality types; resulting e.g., in lack of leadership and work direction). We apply a penalty of factor 0.2 to teams that do not have the full DISC personality spectrum and a penalty of factor 0.4 to teams that have more than one worker with of Dominant personality type (D type). We bound team compatibility to a range $[0, 1]$. Finally, the team compatibility function looks as follows:

$$\textbf{Team Compatibility} = \begin{cases} 0.4 + 0.2 \times (n_{per} - 1), & p_D < 2 \\ 0.2 \times (n_{per} - 1), & p_D \geq 2, \end{cases} \qquad (4)$$

where, n_{per} is the number of all unique personality types, and p_D is the number of workers with a **Dominant** personality type within the team.

3.2.2.3. Team Size

The team size is the third factor that affects team quality in our setting. Literature in small groups research (Moreland, 2010) tends to consider that groups of less than three people do not constitute a team, and that the minimum team size is three. The reason, is that dyads are more ephemeral than larger groups, and certain phenomena like majority/minority relations, coalition formation, and group socialization can only be observed in larger groups. At the same time, social theories underscore the importance of also having an upper critical mass for team collaboration, beyond which the collaboration effectiveness diminishes due to coordination costs (Marwell et al., 1988; Kenna and Berche, 2012). In our setting we apply a penalty factor of 0.1 to teamwork size utility for each additional worker above a maximum threshold of team size five and to each worker needed to reach a minimum team size of three. The team size penalty factor is expected to implicitly guide workers in the self-organized and hybrid approaches to form teams that are within an ideal size range between three and five and discourage them to settle for smaller or larger configurations. The team size function is calculated as follows.

$$\textbf{Team size utility} = \begin{cases} \max(0, 1 - 0.1 \times (S_{MIN} - s_{team})), \\ \quad s_{team} < S_{MIN} \\ \max(0, 1 - 0.1 \times (s_{team} - S_{MAX})), \\ \quad s_{team} > S_{MAX} \\ 1, \quad \text{otherwise}, \end{cases} \qquad (5)$$

where s_{team} is the size of this team, and $S_{MIN} = 3$ and $S_{MAX} = 5$ is the minimal and maximal non-penalized size of a team, respectively.

3.3. Work Coordination Models

We distinguish and compare three work coordination models.

1. The first is a **top-down** model, where the state-of-the-art team formation algorithm Hive appoints teammates without any input from the workers. This strategy approaches TFPs in a controlled, directed, and centralized way. For this model we use the Hive algorithm (Salehi and Bernstein, 2018) designed to optimize team formation from a community-based, top-down approach.
2. The second is a **self-organized** model, where workers govern the team formation processes (grouping and dismantling), with certain rules concerning whether teams should dismantle in the event of minority dissent or not. This approach is inspired by the SOT framework (Lykourentzou et al., 2021) honoring workers' preferences of teammates through a voting system combined with a graph cutting algorithm. We foresee two SOT models called **Radical SOT (R-SOT)** and **Conservative SOT (C-SOT)**. While these two systems share the same bottom-up team formation principles, they differ in the way they handle team cohesion after changes in workers' preferences. Where R-SOT dismantles teams and constructs new ones each time a team member leaves (hence it radically changes team structures), C-SOT preserves teams by retaining their structure and allowing members to leave and join (thus conserving team states where possible). We chose to model two kinds of bottom-up strategies given that certain tasks favor one model over the other (for example radical vs. incremental innovation).
3. The third is a **hybrid** model; this is a mix of top-down and bottom-up team formation strategies where algorithmic intervention supports and is driven by worker feedback. In the hybrid model, network efficiency, tie strength, and workers' agency are combined into a unified system where teams are regularly dismantled in the event that at least one teammate wishes to leave.

3.3.1. Top-Down Model: Hive

For the implementation of the top-down team formation strategy, we adopt Hive (Salehi and Bernstein, 2018), a crowdsourcing collaborative hierarchical team formation model for which community structures dictate network changes. Hive models workers as part of a collaboration graph, with workers as the nodes and the edges corresponding to prior worker collaborations. The objective of the algorithm is to regularly shuffle teams so as to bring together workers with different viewpoints (i.e., far away in the graph), while conserving tie strength. To do so, Hive groups people in teams with one fixed leader, and then intermixes the teams by rotating the people who are not leader. The original Hive paper does not specify how each leader is appointed, or which is the optimal team size to be used. To be able to apply Hive on our setting, we needed to make a decision concerning these two parameters; in both cases we made the decision that is the most favorable for Hive. Concerning team size, we used teams of five. This is the minimum team size for a worker team to have chances to cover all DISC personalities, plus one for the fixed team leader. This way, the Hive teams always have a leader and always have a chance to cover all DISC personality types, i.e., they have a chance to be optimal. Concerning leadership, we appoint the fixed (non-movable) leader of each team to be the team member who has a D personality type, if one such member exists. This way the Hive teams avoid being leaderless, which would result in less-than-optimal results.

In the event of too many workers of personality type D within the worker population, we randomly draw a subset of D-leaders equal to the number of teams. After all team leaders are assigned to their teams, we randomly match workers to the teams, in the same way that Hive randomly initializes the movable team members in the beginning of the task. With these modeling decisions in place, we proceed to model the Hive approach for our simulation. We first introduce the concepts and calculations of network efficiency and tie strength, which are central to the Hive algorithm. We then implement these metrics as part of Hive's objective function, and finally we describe the modeling of the stochastic search algorithm used to find possible team formation moves.

1. **Network efficiency**: The efficiency of a network describes how effectively it transports information across its nodes (Latora and Marchiori, 2001). Network efficiency is usually calculated as the average of the inverse of the minimal path length between every two nodes. By applying network efficiency to the simulation, we attribute the value 1 for all familiar ties (meaning ties linking workers who have collaborated in the past) and the value $+\infty$ to those ties that do not share direct collaborative history. Formally, the network efficiency *NE* in the system is calculated as follows:

$$\mathbf{NE}(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}, \tag{6}$$

   where N is the number of workers in the system and d_{ij} is the minimal path length between node (worker) i and node (worker) j.
2. **Tie strength**: Tie strength represents the level of closeness or affinity between two nodes of a network. In the simulation, tie strength is intended as the calculation of relationships between workers, and ties between nodes represent the workers' collaboration history. Following the Hive computation of tie strength, we apply a logistic function and dampening factor to represent incremental familiarly and progressive detachment, respectively.

Algorithm 1: Stochastic search algorithm. The algorithm attempts to add as many valid rotations as allowed to the network graph, as long as the new rotation surpasses the current state of the objective function (Equation 7) and until either all moves are exhausted or a local maximum is reached (Salehi and Bernstein, 2018).

```
Data: Network graph $G_{odd}$
Result: Network rotation solution
$solution \leftarrow \{\}$;
$bad\_moves \leftarrow \{\}$;
while true do
    $candidate, new\_move \leftarrow$ AddValidMove($solution$, $bad\_moves$);
    if $candidate$ is None then
        return $solution$;
    end
    $G' \leftarrow$ Transform($G_{old}$, $candidate$);
    $G \leftarrow$ Transform($G_{old}$, $solution$);
    if $f(G') > f(G)$ then
        $solution \leftarrow candidate$;
    else
        UpdateBadMoves($bad\_moves$, $new\_move$);
    end
    if random()$\leq \epsilon$ then
        return $solution$;
    end
end
```

   (a) **The logistic function** takes two parameters $k = 8$ and $x_0 = 0.2$ used to simulate the rapid strengthening of relationships at the start of new collaborations (where tie strength is lower) and their slow increment over time.
   (b) **The dampening factor** captures the weakening of tie strength when workers no longer collaborate and are therefore not directly exposed to one another. For its calculation, we adopt the same value as Salehi and Bernstein (2018) ($\lambda = 0.8$).
3. **Objective function**: The objective function of Hive consists of combining network efficiency and tie strength; this is since, in the event of workers changing teams, network efficiency grows as new collaborations emerge (and information gets transported across the network) while tie strength decreases as there are less close relationships. We factor these parameters in the simulated model with a constant value $\alpha = 0.5$ as described in Equation (7). Here, we normalize tie strength by a constant value $c = 0.005$.

$$\mathbf{f(G)} = \alpha \times \text{TieStrength}(G) + (1-\alpha) \times \text{NetworkEfficiency}(G) \tag{7}$$

4. **Stochastic search**: As also discussed by the makers of Hive (Salehi and Bernstein, 2018), effectively rotating teams in order to reach optimality is an extremely complex and non-uni-modular task $[O(2^N)]$. We implement the stochastic search algorithm of Hive as described in the stochastic phase

Algorithm 2: Add valid move algorithm. This algorithm loops for every team and for every worker (that are not team leaders), until it finds a worker and a team (represented by its leader) to meet the following five conditions: 1. the worker is not in the team; 2. current team size is within the system constraints; 3. target team size is within system constraints; 4. this combination is not a **bad move**; 5. this combination is new.

```
Input: Current solution solution, bad moves bad_moves
Result: Candidate solution after adding one move, one valid
        move
leader_ids ← Shuffle(GetTeamLeaders());
all_teams ← Shuffle(GetTeams(solution));
for team t in all_teams do
    for worker w in t do
        if w NOT in leader_ids then
            for leader l in leader_ids do
                if l NOT in t AND
                  Size(t) > $S_{MIN}$ AND
                  Size(l) < $S_{MAX}$ AND
                  w, l NOT in bad_movesAND
                  solution(w) $\neq$ l then
                    candidate, new_move ←
                    AddOneMove(solution, w, l);
                    return candidate, new_move;
                end
            end
        end
    end
end
```

1 (**Algorithm 1**) and phase 2 (**Algorithm 2**). In essence, the stochastic search algorithm finds a random valid move, i.e., it identifies which worker should move to which team, which carries greater utility than the previous move considered by the algorithm. It returns a solution when the search space has been exhausted or if the ϵ value is reached indicating the probability of stopping the search.

3.3.2. Bottom-Up Model: SOT

In the bottom-up model, we simulate team formation on the basis of workers' preferences and affinities. In this context, teams strictly depend on what workers prefer and how likely they are to form effective teams with regards to their personality, knowledge domain, and team size. The simulation represents an abstraction of workers' behavior, performance, and constraints while they form teams in a self-organized manner. During each round workers are allowed to change teams after a deciding and searching phase.

1. **Deciding phase:** In this phase, workers evaluate the strength of their risk appetite against the reward they received in the previous round. The factor with the highest score (being it either risk appetite or reward Equation 1) determines whether that worker will decide to remain in the same team in the following round or whether they will join another team. A higher risk appetite stirs workers to leave and seek new coalitions in search for higher future rewards, whilst a lower risk appetite means that the worker will stay with their existing team even for lower rewards.
2. **Searching phase** Workers who decide to change teams proceed with the search phase, where they perform an evaluation of compatibility of the teammates and teams available to them. Specifically, during this phase, workers assess all possible combinations of teams of four by evaluating three other available workers based on a cosine similarity score of the four manifest attributes, i.e., the attributes of their co-workers that they can readily see (knowledge domain, nationality, educational level, and age). The cosine similarity score does not factor in the average past reward of the workers as it only deals with their manifest profiling attributes. However, in the event that two workers have the same similarity score, their average past reward is considered as a tie breaker. The search phase is further differentiated between the two bottom-up model variations as described below.

    (a) **Conservative SOT (C-SOT)** According to the conservative strategy, existing teams are given priority in choosing whether to admit new members or not. The C-SOT strategy considers existing teams as those that have worked together in the previous round and have at least two team members who decided to continue working together, during the deciding phase. For the rest of the workers and teams that do not fit into this description, the strategy considers these workers as available and unassigned entities. Then, the decision-making process is based on the homophily score (how similar the candidate team members are) constrained by a threshold (Section 3.4) determining the minimum similarity required to form matches. Teams recruit (are matched to) workers who have a similarity score higher than the threshold and higher than the rest of the available workers. If the similarity between existing teams and available solo workers is below the threshold (thus candidate teammates do not classify as sufficiently similar to any given team), the C-SOT model ignores the previously formed teams and matches available teammates based on the highest homophily score. The strategy then puts similar and available teammates into teams of four. In the case of equal similarity between workers or between workers and teams, the strategy prioritizes matches of the highest teamwork quality. Finally, in case that workers still cannot be matched, the C-SOT strategy puts those workers on hold until the next searching phase.

    (b) **Radical SOT (R-SOT)** While the C-SOT strategy attempts to preserve the existing teams' structures even though one or more members decide to leave, in the R-SOT strategy, a team is considered dismantled and all of its members are made available even if one worker from that team decides to leave. This means that available workers have higher chances of forming new teams since they are given access to more options. Besides this difference in the way of handling team deconstruction, the R-SOT follows the same approach as the C-SOT. It too assesses all possible combinations of similarities between four available workers and forms teams of the highest similarity score. In the event that no three workers are considered sufficiently similar to be matched, the R-SOT strategy strives to match workers with existing teams (intended as those that did not lose teammates in the deciding phase). If workers can still not be matched neither with a newly formed nor with an existing team, the team formation model leaves these workers on hold until the next searching phase.

3.3.3. Top-Down and Bottom-Up Models Combined: HiveHybrid

Although bottom-up approaches to crowd TFPs—such as the SOT model—have certain advantages over top-down algorithmic solutions, their spontaneous nature and weak controllability can result in suboptimal solutions. Workers often cannot access the full array of options at once, mostly due to external constraints such as budget, availability, and time. More so, a system that fully relies on the workers' choices to form teams is susceptible to errors of judgment as workers evaluate others subjectively and cannot possess the same global overview of a centralized system. This means that workers cannot always judge the optimality of a match on the basis of both local and global objectives as their angle of vision is often restricted by what they can experience. This locality issue is even more present when the pool of workers is considerably large and workers are limited by how many people they can meet. Under the light of these inherent limitations of fully bottom-up solutions to crowdsourcing TFPs, we also model a blended approach inspired by Prokopenko (2009) who point that self-organization can (and should) be guided by algorithmic top-down mediation. Similar works (Lykourentzou et al., 2010, 2019; Martius and Herrmann, 2012; Nurzaman et al., 2014; Jarrahi et al., 2020)—either through conceptualization or real-life implementations—have proposed **guided self-organization** as the ideal strategy linking worker agency with algorithmic optimization. Our implementation of guided self-organization differs in the way it is applied to a simulated collaborative crowdsourcing scenario where workers are recommended by the algorithm whether to change teams or not. The HiveHybrid model is designed precisely to combine global objectives with local constraints in large-scale collaborative crowdsourcing. The system combines a bottom-up worker-centric SOT model with a top-down community-based Hive model. In the HiveHybrid, workers are allowed to decide to leave a team or remain as their choice is honored and optimized through a community-based team rotating algorithm. The algorithm identifies possible moves (rotations that would benefit the global objective function) and the workers can either accept or reject this offer if their appetite for exploration (risk appetite) indicates so.

3.4. Experimental Parameterization

Our simulation is designed to run a series of experiments where different populations and team formation models are tested and evaluated for their best, worst, and average teamwork quality. The following are the experimental parameters and corresponding settings used for this study. For the implementation of the Hive algorithm both as a baseline for top-down allocation and as part of the HiveHybrid model, we use the same parameters stated in the work by Salehi and Bernstein (2018).

1. **Experiment rounds (n)**: By rounds we intend the collaboration cycles during which workers form teams and collaborate. For this study, we used a fixed experiment of 10 rounds.
2. **Teamwork quality**:We calculate teamwork quality as follows. We first generate a batch of user agents as described in Section 3.2. For this batch, we run the simulation six times, each time extracting the best, average, and worst teamwork values, and then calculating the mean of those values to get the best, average, and worst teamwork quality of the batch. We repeat the process for thirty independent batch runs and average out the results. The procedure is designed to smooth out random fluctuations and yield less noisy simulation results.
3. **Population (x)**: The default population size is set to 20 workers. We consider this to be a rounded estimation of a basic size of participation required for creative tasks of this kind (online hackathon, expert crowdsourcing collaboration, etc.). Then, to examine generalizability, we gradually increase this number and experiment with larger populations ($[30, 40, 50, \dots, 100]$).
4. **Team size threshold (S_{MIN}, S_{MAX})**: We constrain teams within a range of three (minimal size) and five (maximal size) teammates. We apply these threshold since we expect smaller teams to be hindered by a shortage of knowledge domains and personalities while larger teams to be taxed by coordination and communication costs, as explained in Section 3.2.2.3.
5. **Risk appetite (β)**: We represent worker's risk appetite using two mirror symmetric distributions. For the explorative behavior (high risk appetite) we use a beta distribution of negative parameter range ($\beta \in [-5, -2]$), while for the exploitative behavior (low risk appetite) we use its symmetric positive parameter range ($\beta \in [2, 5]$). We further model a neutral risk level to be bounded within a probability distribution of $\beta = 2$.
6. **Homophily threshold (θ)**: The homophily threshold determines the extent to which people are willing to accept working with others based on their in-between attribute cosine similarity. Since we use four dimensions to determine workers' similarity (knowledge domain, nationality, educational level, age), we bound the homophily threshold within the range $[1, 4]$. The workers' default homophily threshold is set to 2.8 meaning that any similarity below this value is not considered sufficient to form a match.
7. **Teamwork quality coefficients (π, μ)**: The coefficients π and μ represent the weights attributed to team skill and interpersonal compatibility respectively. The default values are set to 0.4 for both π and μ. While we use these coefficients to adjust the weights of team skill and interpersonal compatibility, the same weight is taken off from the team size (1 - $\pi - \mu)^2$.

[2] This applies less weight to that factor.

4. RESULTS

In Section 4.1, we compare teamwork quality across the four models: Hive, C-SOT, R-SOT, and HiveHybrid and address the first research question (**RQ1: How does bottom-up team formation compare with top-down and hybrid approaches?**). Next, we address the second research question (**RQ2: How do population behavioral tendencies affect the outcome of bottom-up online teamwork?**) in three Sections, one for each of RQ2 sub question: Sections 4.2 and 4.3 examine teamwork quality according to changes in the workers' risk appetite and population size distributions, respectively, while The descriptive statistics report the mean and standard deviation (sd) of the model's teamwork quality. The standard deviation indicates the average amount of variability within a set of experiments. For example, *mean* = 0.716 and *sd* = 0.024 of the R-SOT model's best quality indicate, respectively, the mean and the standard deviation of the best teamwork gathered from thirty independent batch runs, as explained in the Methodology (Section 3.4).

4.1. Comparing Models: Radical Bottom-Up Yields the Highest and Lowest Teamwork Quality

In this Section we address the first research question, namely **RQ1: How does bottom-up team formation compare with top-down and hybrid approaches? Figures 2–5** shows the results of running a comparative study with all four models (R-SOT, C-SOT, Hive, and HiveHybrid) and utilizing the parameters stated in Section 3.4. We analyse the results below.

4.1.1. Best Teamwork Quality

R-SOT has the highest average best quality (*mean=0.716*, *sd=0.024*), followed by C-SOT (*mean = 0.698*, *sd = 0.022*), HiveHybrid (*mean = 0.689*, *sd = 0.030*), and Hive (*mean = 0.683*, *sd = 0.029*) indicating that bottom-up models outperform the rest in forming the most competitive teams. Although standard deviations are relatively close across models, the standard error is greater in HiveHybrid than all other models possibly due to the unpredictability of combining suggested changes from the top-down community-based model with workers' decision.

4.1.2. Average Teamwork Quality

R-SOT still performs better than the rest, although its mean is only marginally higher than the other models (*mean = 0.572*, *sd = 0.016*) followed by HiveHybrid (*mean = 0.572*, *sd = 0.023*), Hive (*mean = 0.567*, *sd = 0.021*), and C-SOT (*mean = 0.563*, *sd = 0.018*). In this comparison analysis, standard deviations are fairly close, while the standard error of HiveHybrid remains, by far, the largest in this comparison of the average teamwork quality. In fact, HiveHybrid's large standard error is present in all evaluations of teamwork quality.

4.1.3. Worst Teamwork Quality

HiveHybrid has the least worst teamwork quality as its mean is above all others (*mean = 0.467*, *sd = 0.024*), followed by Hive (*mean = 0.460*, *sd = 0.019*), C-SOT (*mean=0.429*, *sd = 0.018*), and R-SOT (*mean = 0.416*, *sd = 0.020*). These final results indicate that the hybrid model is efficient at reducing the segregating patterns present in bottom-up systems, which lead to great variations of teamwork quality. Although the standard deviations are fairly close across models, HiveHybrid retains the largest standard error making it less consistent in its team formation.

4.1.4. Statistical Analysis: R-SOT Outperforms in the Best and Loses at the Worst Teamwork Quality

Running a one-way ANOVA test on the results from the comparison of the teamwork quality between the four models we find the following.

- **Best teamwork quality:** The best teamwork quality is statistically significant between groups [$F_{(3,\ 116)} = 10.477, p < 0.001$]. Specifically, R-SOT performed significantly better than C-SOT (*p*=0.003), Hive (*p*<0.001), and hybrid Hive (*p*=0.001).
- **Average teamwork quality:** No statistical difference was found between models when comparing their average teamwork qualities [$F_{(3,\ 116)} = 1.394, p < 0.248$].
- **Worst teamwork quality:** We found statistically significant results between groups with the Worst teamwork quality. [$F_{(3,116)} = 35.122, p < 0.001$]. Here, R-SOT performed significantly worst than C-SOT ($p = 0.036$), Hive ($p < 0.001$), and hybrid Hive ($p < 0.001$). The R-SOT model also performed significantly poorly compared to Hive ($p < 0.001$), and hybrid Hive ($p < 0.001$). Lastly, Hive and hybrid Hive did not differ significantly.

4.2. High Levels of Risk Appetite Segregate Teamwork Quality in Bottom-Up Models

This Section, combined with the two Sections that follow, dives deeper into the performance of bottom-up large-scale collaboration, and contributes to answering the second research question **RQ2: How do population behavioral tendencies affect the outcome of bottom-up online teamwork?** Specifically, it deals with the sub-question **RQ2.1: How do different risk appetites affect teamwork output in bottom-up models?** Looking at the highest teamwork quality (**Figure 3**) results across the two bottom-up models (R-SOT and C-SOT) and controlling for levels of the β value of the risk appetite distribution in the population, we note the following.

4.2.1. Best Teamwork Quality

The radical self-organization approach (R-SOT) achieves better results in terms of the best teamwork quality ($mean = 0.711, std = 7.42e{-3}$) compared to the conservative self-organized approach (C-SOT) ($mean = 0.691, std = 8.99e{-3}$). This result indicates that the overall risk level of a population directly affects the workers' chances of forming optimal teams in bottom-up team formation strategies. Furthermore, high levels of risk appetite within a crowd population seem to be particularly beneficial to systems advocating radical changes in team structure. By lowering the overall risk appetite levels in both R-SOT and C-SOT, the performance of the best teamwork quality progressively suffers, dropping from 0.72 to 0.70 for R-SOT, and from 0.71 to 0.68 for C-SOT.

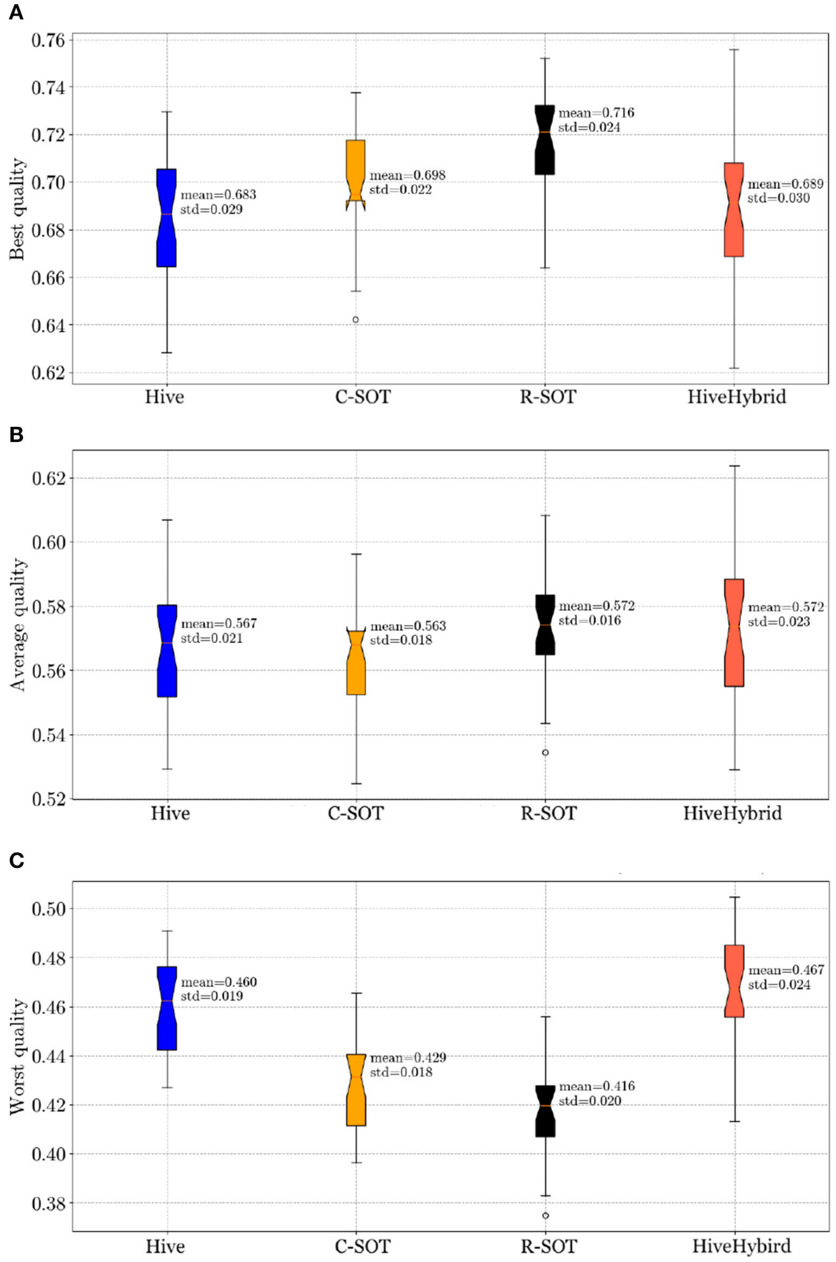


FIGURE 2 | Teamwork quality comparison across four models: Hive, C-SOT, R-SOT, HiveHybrid. The boxplot displays the mean, standard deviation, and standard error of the teamwork quality. Overall, the best teamwork quality (Equation 2) belongs to the bottom-up models R-SOT (mean = 0.716) and C-SOT (mean = 0.698) followed by hybrid (mean = 0.689) and top-down (mean = 0.683). The average performance is fairly equal between models, with HiveHybrid and R-SOT having a slightly higher mean (mean = 0.572). The worst teamwork quality comes from the bottom-up models (R-SOT mean = 0.416, C-SOT mean = 0.429), followed by Hive (mean = 0.460). HiveHybrid performs the best at forming the least worst teamwork quality (mean = 0.467). **(A)** Best teamwork quality for Hive, C-SOT, R-SOT, and HiveHybrid. **(B)** Average teamwork quality for Hive, C-SOT, R-SOT, and HiveHybrid. **(C)** Worst teamwork quality for Hive, C-SOT, R-SOT, and HiveHybrid.

4.2.2. Average Teamwork Quality

Results for the average teamwork quality are similar across both bottom-up models with R-SOT ($mean = 0.569, std = 2.16e{-3}$) and C-SOT ($mean = 0.562, std = 3.03e{-3}$) sharing similar outputs. These results indicate that the risk levels do not necessarily affect average performance despite of which bottom-up team formation strategy is used.

4.2.3. Worst Teamwork Quality

Although the radical self-organized approach performed the highest when it came to the best teamwork quality, we observe that this approach is also the one that performs the worst from the two models R-SOT ($mean = 0.422, std = 2.90e{-3}$) compared to C-SOT ($mean = 0.435, std = 2.43e{-3}$). This result indicates that radically bottom-up approaches may inadvertently

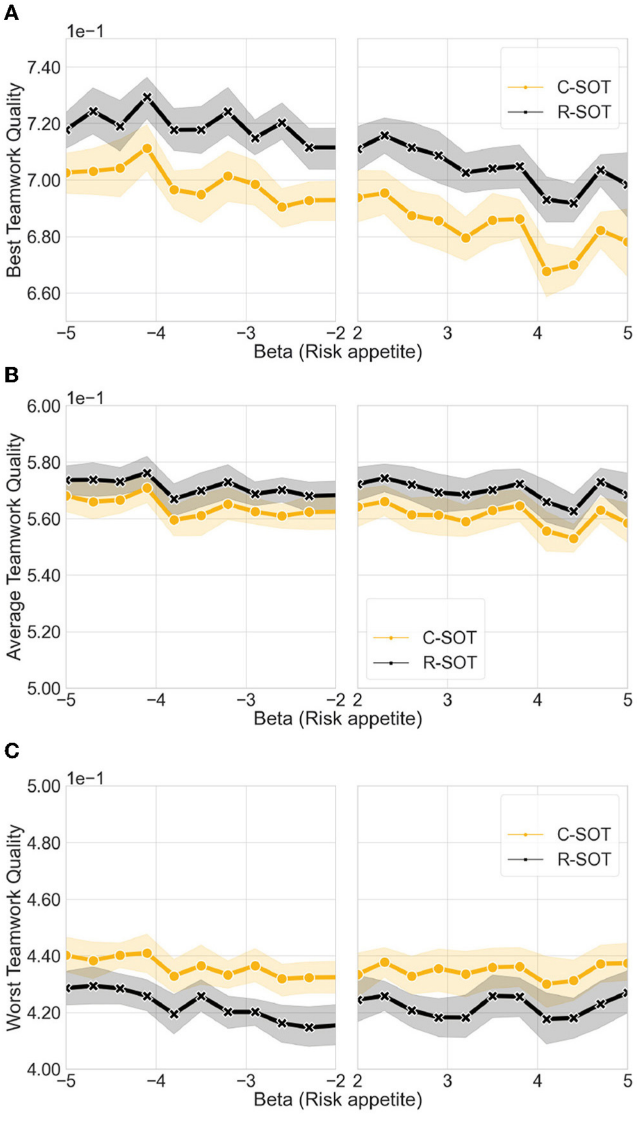


FIGURE 3 | Comparison of the best and worst teamwork quality between bottom-up models, namely C-SOT and R-SOT, with **varying risk appetite levels**. The x axis illustrates the different risk levels generated according to two mirroring beta distributions: negative values ($\beta \in [-5, -2]$) illustrate an exploratory, risk-prone user behavior (the more negative the more risk-prone); positive values ($\beta \in [2, 5]$) illustrate an exploitative, risk-averse behavior (the more positive the more risk-averse). We observe that the best teamwork quality is affected by risk appetite, and that it decreases for both models as the users' willingness to change teams decreases. The average and worst teamwork quality remain unaffected by changes in the user population's risk levels. **(A)** Best teamwork quality for C-SOT and R-SOT with different risk appetite levels. **(B)** Average teamwork quality for C-SOT and R-SOT with different risk appetite levels. **(C)** Worst teamwork quality for C-SOT and R-SOT with different risk appetite levels.

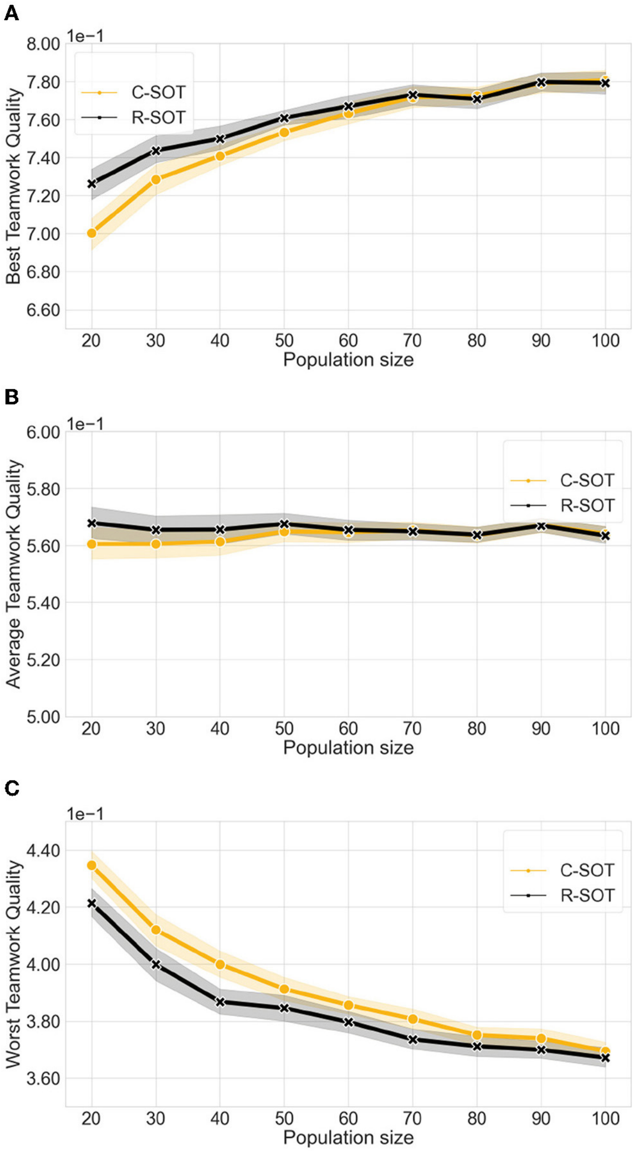


FIGURE 4 | Comparison of the best and worst teamwork quality between two bottom-up models (C-SOT and R-SOT) with different population sizes. The x axes show the different simulated population sizes per hackathon in the $\in [20, 100]$ range. We observe that the best teamwork quality for both bottom-up models improves as the population grows from 20 to 90 individuals, and workers have more choice of teammates, reaching stability with populations of more than 90 and maintaining a best teamwork quality of ≈ 0.77 in both models. However, the worse teamwork quality also decreases steadily in both models lowering from ≈ 0.43 to ≈ 0.37 as the population grows indicating that large populations are not always beneficial to the performance of all teams. **(A) Best teamwork quality** for C-SOT and R-SOT with different **population sizes**. **(B) Average teamwork quality** for C-SOT and R-SOT with different **population sizes**. **(C) Worst teamwork quality** for C-SOT and R-SOT with different **population sizes**.

exacerbate the differences between teams, with the best workers choosing to team up with the best workers, leaving many of the average or low-performing workers behind, and causing segregated quality outputs.

4.3. Large Populations Strengthen Strong Teams in Bottom-Up Models

This Section answers the sub-question: **RQ2.2: How do different population sizes affect teamwork output in bottom-up**

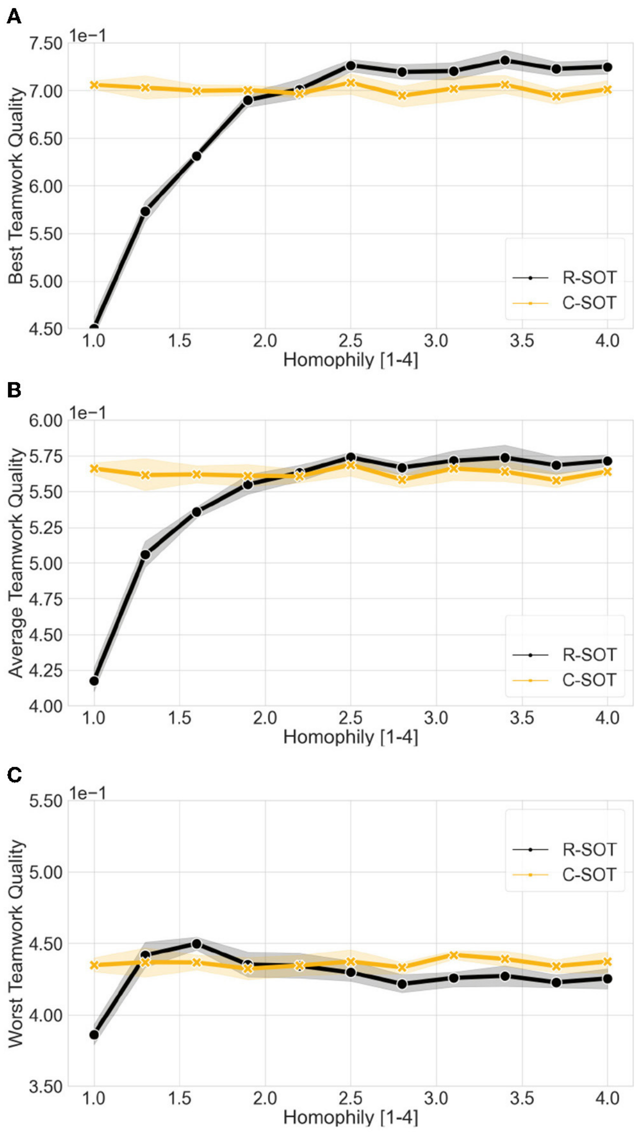


FIGURE 5 | Comparison of the best and worst teamwork quality between bottom-up models (C-SOT and R-SOT) with different homophily thresholds. The x axes show the workers' different homophily levels ($\theta \in [1, 4]$) where the lowest θ value represents the highest possible tolerance toward differences between workers' attributes, and the highest θ value the lowest tolerance. Even though the best teamwork quality for the two models improves as the homophily threshold grows, particularly with thresholds $\theta > 2$, the worst teamwork quality remains overall stable with the exception of R-SOT peaking around a threshold $\theta \approx 1.5$ before settling around a worst teamwork quality of 0.42 with $\theta > 2.5$. **(A) Best teamwork quality** for C-SOT and R-SOT with different **homophily (preference of working with similar teammates) thresholds ($\theta \in [1, 4]$)**. **(B) Average teamwork quality** for C-SOT and R-SOT with different **homophily thresholds ($\theta \in [1, 4]$)**. **(C) Worst teamwork quality** for C-SOT and R-SOT with different **homophily (preference of working with similar teammates) thresholds ($\theta \in [1, 4]$)**.

models? In the basic experimental setting (Section 3.4), we used a population size of 20 workers, which is a size that guarantees that workers can process the information concerning all other candidate co-workers effectively. However, the population size is a factor that can critically affect performance, as it is known to affect the worker collective's coordination costs. A larger population means a larger search space of available candidate teammates, and therefore more effort needed by the workers to process suitable teammates (Kittur and Kraut, 2008). We simulate nine separate and increasing population sizes starting from 20 (our basic simulation setting) and going up to 100 workers per pool to observe how the average best, worse, and median teamwork quality vary accordingly.

4.3.1. Best Teamwork Quality

With an incremental growth in population size, both R-SOT and C-SOT improve their best performance shifting from an average best teamwork quality of 0.70 to one of 0.78. This result shows that bottom-up approaches particularly benefit from large scale participation as they rely on the diversity of workers' backgrounds and skills to form optimal teams.

4.3.2. Average Teamwork Quality

As also observed in the previous Section for the parameter of risk appetite, we observe that the average teamwork quality neither benefits nor deteriorates from changes in population size and it remains relatively constant around 0.568.

4.3.3. Worst Teamwork Quality

Lastly, the worst quality of bottom-up approaches drops from an average of 0.43 to an average of 0.36. The worst quality of R-SOT (*mean*=0.416) is indeed worse than C-SOT (*mean*= 0.429). This result may be explained by the fact that the R-SOT strategy dismantles teams having at least one unsatisfied worker and gives access to many more available workers of attractive attributes who can therefore settle for higher payoffs in the next round. Similarly to the results of the previous Section, here too we observe that the radical model (R-SOT) is the one yielding the highest best and the lowest worst quality.

4.4. Similar Workers Produce Higher Teamwork Quality in Bottom-Up Models

In this section, we address the last sub-question **RQ2.3: How does homophily, i.e., the tendency to prefer working with similar teammates, affect teamwork output in bottom-up models?** The homophily threshold determines the workers' tolerance to the diversity of attributes in others. Setting low homophily thresholds allows workers in the simulation to form larger teams since they are more open toward diverse team members. Using higher homophily thresholds pushes workers to carefully choose their teammates and only be interested in those who are most similar. Since there are four similarity attributes in the calculation of the cosine similarity score (knowledge domain, nationality, educational level, age), we use a homophily range of $[1, 4]$ with a step of 0.3 allowing us to test ten variations of the threshold search space.

4.4.1. Best Teamwork Quality

Testing all simulated thresholds we observe that the best teamwork quality of C-SOT is not affected by changes in homophily. However, R-SOT's best teamwork quality rapidly grows as the threshold increases from 0.50 with $\theta = 1.0$ to 0.70 with $\theta = 2.8$. After this growth, the best teamwork quality of the R-SOT model stabilizes and does not improve.

4.4.2. Average Teamwork Quality

Similarly, the average teamwork quality is not significantly affected by changes in homophily in the C-SOT model while the R-SOT's average teamwork quality grows quite rapidly from 0.44 to 0.57 (from $\theta = 1$ to $\theta = 1.5$), and continues to rise before stabilizing around 0.72 with $\theta > 2$.

4.4.3. Worst Teamwork Quality

The worst teamwork quality is not affected by changes in homophily threshold for the C-SOT with the worst teamwork quality remaining stable at around 0.44. R-SOT's worst teamwork quality is more drastically affected by an increase in the homophily threshold, with an increase in quality between $\theta = 1.0$ and $\theta = 1.6$ and then a gradual decrease and stabilization after $\theta = 2.5$. This is a similar pattern (sharp rise and then stabilization) like the one we saw R-SOT following in the best and average teamwork quality results, albeit with less intensity as we go from best to worst quality.

5. DISCUSSION

5.1. Bottom-Up Models Are Ideal for Large Scale Crowdsourcing Collaborative Innovation

We observe that the bottom-up models R-SOT and C-SOT are more effective than the top-down (Hive) and hybrid solutions (HiveHybrid) at forming teams with the best teamwork quality. In these self-organizing systems, workers seek collaborators based on how likely they are to explore the search space and how tolerant they are toward diverse teammates. Whether workers will explore further teammates depends on the reward they received with their old team, bounded by their risk appetite. For example, low-risk workers will keep working with the same teammate even if they did not get a high reward, while high-risk workers will change more frequently. Although skill is therefore not explicitly present in the worker's search function, since it is a latent feature that workers cannot directly have access to concerning their teammates, we observe that *gradually workers discover their in-between skills by implicitly evaluating the results of their existing collaborations against others through the rewards each team received.*

This result shows that *bottom-up systems lift the requirement for intensive skill profiling before they can make good team formation decisions*, and it is important for platforms for multiple reasons. First, it renders bottom-up models *more appropriate for innovation-related tasks* for which the exact skills that will be needed to solve the task are not easily measurable or even known a priori to the collaborative platform (Gerber et al., 1999). Second, by lifting the requirement for designing tailor-made profiling tests and team formation algorithms, *bottom-up models are more cost-effective than their top-down or hybrid counterparts*, and are also particularly useful for introducing new tasks in the platform for which no profile information concerning worker competencies or matching mechanism is yet present. One important point to make here is that these advantages of bottom-up models refer to the best teamwork quality achieved, but that, at the same time, these models also tend to segregate the worst-performing teams. In other words, *bottom-up models help form principally strong and competitive coalitions which may form at the expense of other teams*. Therefore, self-organization could be more appropriate and cost-effective for commercial platforms targeting at the competitive production of tasks, rather than tasks for which "no worker is left behind" (e.g., in educational settings). Overall, our comparative analysis of team formation system models in crowdsourcing collaborative innovation indicates that platforms can "trust the crowd" to form teams as long as they favor competition over cooperation and as long as they prefer competitive teams over a centralized re-distribution of social capital.

5.2. Hybrid Systems Are Best for Semi-centralized Social Capital Redistribution

As we have seen, bottom-up systems are the best at producing teams of the highest teamwork quality. However, having the best team is not the sole representative metric of collective performance in social cooperative scenarios. In fact, in the case of our bottom-up models, the worse-off teams do not seem to benefit from a self-organizing system as their teamwork quality notably suffers compared to the worst teams from algorithmic-mediated team formation solutions. Especially, *the hybrid approach HiveHybrid is the most effective at bridging the gap between best and worst teamwork qualities* forming teams that are closer in the way they perform. This ability to redistribute resources among a population to help all teams achieve similar teamwork quality is exceptionally favorable in settings where global objectives are of equal importance to local interactions as it is in the case of groups of learners. Massive Open Online Courses (MOOCs) are an example of large-scale collaborative settings that would benefit from fair semi-centralized clustering to allow all learners to partake in useful and educational teamwork. It could also help with reducing drop-outs as learners would be motivated to partake in teams from which they can learn and share knowledge. Through hybrid approaches to team formation, online learners would be given the final decision over algorithmic prompts carrying decisive advantages over fully top-down implementations that typically disregard individual preferences.

Control of social capital and resource redistribution is not the only advantage of hybrid systems. Connecting workers' decisions—which are by definition local and discrete—with global utility and centralized coordination *could help with situations where workers can no longer process information by themselves*. Often with online crowdsourcing innovation projects, several hundred individuals take part in events and seek collaborators. As much as the team formation system relies on their ability to self-organize to produce optimal teamwork outcomes, there is always the looming risk that workers cannot assess more than a limited amount of possible collaborators at a time. This means that workers are likely to miss out on potentially better matches as they simply cannot have a comprehensive view of all options unless they scout the entire pool of workers. However, with prolonged exploration workers do not have the time to settle into teams and may discard ideal collaborators to continue their search. Due to this extraneous cognitive overload and excessive search space, *a system that can fittingly combine algorithmic mediation and worker agency—such as HiveHybrid—could be a more suitable alternative to fully self-organized and fully top-down systems even though they may be less effective at forming highly competitive teams*.

5.3. Generalizing to Other Collaborative Settings

In this study, we simulated an open collaboration scenario where crowds gather to collaborate on a complex problem. We chose a hackathon as an example of a design-sprint-like event for which crowds compete for prizes while collaborating in teams. As in reality, our simulations represent workers forming project teams either through top-down mediation (upon organizers' decision) or bottom-up negotiation (workers choose their teammates and self-organize). Although traditionally confined to software development, hackathons have developed to serve other scopes, for example, by hosting charity events, public memorials, professional networking, and more, and are therefore much broader than their cryptography development ancestor (Briscoe, 2014). For this reason, hackathons can be used as general-purpose initiatives to attract crowd participation and gather expertise and innovation. Online hackathons have also become attractive mediums for the involvement of citizen crowds in decision-making processes (Temiz, 2021). For example, "Hack The Crisis" is, to date, the most popular crowdsourced global movement connecting crowds to solve complex societal challenges such as pandemics prevention and emergency response (Hack the Crisis Team, 2021). The chosen setting could be applied to large-scale crowd empowerment through open challenges, open education, and social impact.

Regarding the components of the simulation, we modeled only an abstract set of worker skills especially since some hackathons' organizers filter attendance based on functional background and expertise with the intent to harvest specialized knowledge from the crowd. However, the worker model could be easily expanded to other tasks and settings. For example, in a scenario where students form teams, their attributes would represent interests, preferences, and abilities instead of the functional background, personality, and skill as we modeled in this study. Furthermore, some hackathons are characterized by rounds of sprints which we have devoted to individual/algorithmic decision-making and search space. Moreover, in real-life hackathons, it is not unusual that these rounds provide organizers regular opportunities to monitor and evaluate teamwork as the event unfolds. In our study, we use the same concept to evaluate teamwork quality and to allow workers (and algorithms) to rotate teams. From MOOCs to citizen science, these elements of the system can be adjusted to correspond with periods of recollection and assessment that are often present in large-scale crowdsourcing activities. Finally, hackathons usually end up with a selection of the best projects and the best teams, which, in this case, is the main metric for assessing the adequacy of the team formation system models. Generalizing this setting to other scenarios, we suggest that the evaluation could be adjusted to whichever factor the event organizer wishes to assess (e.g., communication, coordination, the balance of contribution and effort, etc.). For example, teams of learners will be likely evaluated based on mutual support, cohesion, and effort, thus differing from software development teams focused on product quality, team efficiency, and profitability.

6. LIMITATIONS

In this section, we list and discuss four main system design choices that could be improved or modified in future studies as follows.

Modeling worker attributes and recruitment through AMT may not be comparable to other platforms. In this paper, we have used AMT as the platform of reference for modeling worker demographic attributes (Difallah et al., 2015; Lykourentzou et al., 2016a) and recruitment. This choice allowed us to ascertain a degree of reliability and applicability since the demographic distribution adhered to existing statistical records. However, as crowdsourcing platforms evolve and differentiate, many more platforms offer like-minded individuals ways to collaborate and participate in disparate projects. The most prevalent crowd population on platforms facilitating creative tasks, such as OpenIDEO, Upwork, Fiverr, or even creative hubs, may have different demographic attributes than their AMT counterpart, which is mainly used to serve micro-tasks. We strongly encourage future studies to consider additional platforms of reference to model workers' profiles (such as educational level, personality, and age) and recruitment, which could be more relevant to creative hackathons and complex problem-solving.

6.1. Homophily Threshold Modeled on the Entire Population

Unlike worker risk appetite, homophily was modeled on the whole population as a shared threshold rather than on an individual-to-individual case through a non-uniform probability distribution. This modeling choice also means that it is not possible to identify how individual homophily might have affected the behavior of a worker in a pool with diverse homophily attitudes. In future studies, modeling the personal preferences of collaborators would help to fine-grain our assessment of the impact of homophily in team formation. It would also help with evaluating how different attitudes toward diversity combined affect the formation of more or less stable teams and to what extent it influences teamwork quality.

6.2. Risk Appetite Goes Beyond the Tendency to Explore Collaborators

In our model of the workers, we attributed risk appetite to the individual tendency to explore novel collaboration. In this context, risk appetite can be thought of as a behavioral property encompassing one's curiosity and extroversion. Nonetheless, risk appetite could also determine one's preference for a particular task and ways of executing it. Modeling task choices, task execution, and effort as part of the workers' risk appetite would also determine their stress and energy levels and delineate a finer-grained representation of human behavior (Chiang et al., 2021). We, therefore, suggest extending the significance and functionality of the risk appetite attribute in future simulations.

6.3. Sensitivity Analysis Limited to Bottom-Up Models

After comparing the four models (C-SOT, R-SOT, Hive, and HiveHybrid) on a set of specific parameters, we have systematically varied the parameters of risk appetite, population size, and homophily threshold for the comparison of the bottom-up models. This analysis permitted us to examine in detail the models' response to varying population behavioral patterns, across the three aforementioned worker attributes. For our main scenario we have chosen a specific and fixed set of parameters; although these modeling choices have been based on the literature, they do limit the applicability of our results to the specific population characteristics. Performing a systematic sensitivity analysis for the main scenario can, in the future, permit to examine whether the current results can be generalized to scenarios with other demographics or whether there are any mixed effects, for instance between the team size and the workers' homophily threshold.

6.4. Teamwork Quality Function Limited in Scope

Our evaluation of teamwork quality is based on the assumption that certain attributes together matter most in determining the probability of success of a team. We identified team skill, interpersonal compatibility, and team size as the determining factors. Although these factors have been shown in the literature to critically affect teamwork performance, there may be additional aspects of the collaboration that also play a significant role depending on the real-world task at-hand. For example, communication quality, the ability to think out-of-the-box as a team may also affect the final result. Follow-up studies could therefore examine additional quality metrics and even evaluate different methods for calculating teamwork quality than the one used in this study.

6.5. Worker Search Space Unhindered by Cognitive and Temporal Constraints

Another assumption present in this study is that workers are not constrained in their search of available teammates. This means that if a large pool of workers is available, workers can evaluate all possible team combinations and pick the one with the highest utility. In real-life settings, this is not always possible as information may be missing and time and resources may be lacking to carry out a thorough evaluation of this kind. Future simulations should take into account the limitations that workers may face when assessing others, especially as the size of the worker pool increases, and convey these in their definition of the search function. It is possible that workers can truly only process a few candidates at a time, and their judgment can be affected by presentation or popularity biases.

6.6. Hive Is One of Many Kinds of Top-Down Models

Our comparison of different team formation models uses only one top-down approach, namely Hive. Although the Hive algorithm is a latest state-of-the-art community-based solution, it does not represent many other kinds of top-down approaches. Due to this limitation, our comparison cannot be entirely generalized to other top-down team formation systems, aside from the acknowledgement that they do not grant worker agency in decision-making. Testing other approaches, such as bi-partite graphs and stable matching algorithms will give future studies more comprehensive knowledge of the effectiveness of these approaches in collaborative crowdsourcing scenarios and how they compare to self-organized and hybrid solutions.

7. CONCLUSION

With the rapid growth of crowdsourcing platforms used for collaborative innovation generation and citizen participation, team formation among members of a crowd becomes increasingly pertinent. This study evaluates how different approaches to crowdsourcing team formation impact teamwork through bottom-up, top-down, and hybrid models. Using a simulated hackathon scenario, we gathered results from the collaboration between strategic worker agents showing that bottom-up models are convincingly more effective at forming highly competitive teams but do not succeed at redistributing equally resources within the crowd population. On the contrary, the hybrid system which combines bottom-up worker agency with top-down algorithmic mediation bridged this gap by forming teams of closer teamwork quality. The purely top-down approach performed averagely whilst still limiting worker agency in team formation. We further observe that high-risk appetite levels, large population sizes, and high homophily thresholds of the involved crowd worker population positively affect teamwork quality in bottom-up approaches. This study furthers our assessment of the impact of self-organization in large-scale collaborative crowd innovation and helps the design of systems incorporating agency in algorithmic mediation in team formation.

DATA AVAILABILITY STATEMENT

The code supporting this article is available as Open Source in: https://github.com/ilykour/self_organised_teams.

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version.

REFERENCES

Ahmed, F., Dickerson, J., and Fuge, M. (2020). Forming diverse teams from sequentially arriving people. *J. Mech. Des*. 142:111401. doi: 10.1115/1.4046998

Ananny, M. (2016). Toward an ethics of algorithms: convening, observation, probability, and timeliness. *Sci. Technol. Hum. Values* 41, 93–117. doi: 10.1177/0162243915606523

Anzola, D., Barbrook-Johnson, P., and Cano, J. I. (2017). Self-organization and social science. *Comput. Math. Organ. Theory* 23, 221–257. doi: 10.1007/s10588-016-9224-2

Avis, D. (1983). A survey of heuristics for the weighted matching problem. *Networks* 13, 475–493. doi: 10.1002/net.3230130404

Barnes, S.-A., Green, A., and De Hoyos, M. (2015). Crowdsourcing and work: individual factors and circumstances influencing employability. *New Technol. Work Employ*. 30, 16–31. doi: 10.1111/ntwe.12043

Berg, J. (2015). Income security in the on-demand economy: findings and policy lessons from a survey of crowdworkers. *Comp. Lab. L. & Pol'y J*. 37, 543.

Berger-Tal, O., Nathan, J., Meron, E., and Saltz, D. (2014). The exploration-exploitation dilemma: a multidisciplinary framework. *PLoS ONE* 9,e95693. doi: 10.1371/journal.pone.0095693

Betts, A., and Bloom, L. (2014). *Humanitarian Innovation: The State of the Art*. United Nations Office for the Coordination of Humanitarian Affairs (OCHA).

Briscoe, G. (2014). *Digital Innovation: The Hackathon Phenomenon*.

Carless, S. A., and De Paola, C. (2000). The measurement of cohesion in work teams. *Small Group Res*. 31, 71–88. doi: 10.1177/104649640003100104

Centola, D., Gonzalez-Avella, J. C., Eguiluz, V. M., and San Miguel, M. (2007). Homophily, cultural drift, and the co-evolution of cultural groups. *J. Confl. Resolut*. 51, 905–929. doi: 10.1177/0022002707307632

Chiang, C.-E., Chen, Y.-C., Lin, F.-Y., Feng, F., Wu, H.-A., Lee, H.-P., et al. (2021). "'I got some free time': investigating task-execution and task-effort metrics in mobile crowdsourcing tasks," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama), 1–14. doi: 10.1145/3411764.3445477

Costa, A. C., Fulmer, C. A., and Anderson, N. R. (2018). Trust in work teams: an integrative review, multilevel model, and future directions. *J. Organ. Behav*. 39, 169–184. doi: 10.1002/job.2213

De Dreu, C. K., and West, M. A. (2001). Minority dissent and team innovation: the importance of participation in decision making. *J. Appl. Psychol*. 86,1191. doi: 10.1037/0021-9010.86.6.1191

De Stefano, V. (2015). The rise of the just-in-time workforce: on-demand work, crowdwork, and labor protection in the gig-economy. *Comp. Lab. L. & Pol'y J*. 37,471. doi: 10.2139/ssrn.2682602

Degli Antoni, G., Fia, M., and Sacconi, L. (2021). Specific investments, cognitive resources, and specialized nature of research production in academic institutions: why shared governance matters for performance. *J. Instit. Econ*. 18, 1–22. doi: 10.1017/S1744137421000655

Deitz, C. (2016). Pragmatism and mechanical Turk: citizenship and labor rights in digital communities of knowledge. *J. Media Ethics* 31, 264–266. doi: 10.1080/23736992.2016.1228816

Difallah, D., Filatova, E., and Ipeirotis, P. (2018). "Demographics and dynamics of mechanical Turk workers," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Los Angeles, CA), 135–143. doi: 10.1145/3159652.3159661

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., and Cudré-Mauroux, P. (2015). "The dynamics of micro-task crowdsourcing: the case of amazon mTurk," in *Proceedings of the 24th International Conference on World Wide Web* (Florence), 238–247. doi: 10.1145/2740908.2744109

Eugene, N., Lee, C., and Famoye, F. (2002). Beta-normal distribution and its applications. *Commun. Stat. Theory Methods* 31, 497–512. doi: 10.1081/STA-120003130

Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Inform. Organ*. 28, 62–70. doi: 10.1016/j.infoandorg.2018.02.005

Florisson, R., and Mandl, I. (2018). "Platform work: Types and implications for work and employment-Literature review," in *European Foundation for the Improvement of Living and Working Conditions*.

Furlow, L. (2000). Job profiling: building a winning team using behavioral assessments. *J. Nurs. Administr*. 30, 107–111. doi: 10.1097/00005110-200003000-00001

Gaikwad, S., Morina, D., Nistala, R., Agarwal, M., Cossette, A., Bhanu, R., et al. (2015). "DAEMO: a self-governed crowdsourcing marketplace," in *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Daegu), 101–102. doi: 10.1145/2815585.2815739

Gaikwad, S. N. S., Whiting, M. E., Gamage, D., Mullings, C. A., Majeti, D., Goyal, S., et al. (2017). "The DAEMO crowdsourcing marketplace," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, OR), 1–4. doi: 10.1145/3022198.3023270

Gerber, C., Siekmann, J., and Vierke, G. (1999). "Holonic multi-agent systems," in *Self-organising Software*. Natural Computing Series, eds G. Di Marzo Serugendo, M. P. Gleizes, and A. Karageorgos (Berlin; Heidelberg: Springer). pp. 238–263.

Gilson, L. L. and Shalley, C. E. (2004). A little creativity goes a long way: an examination of teams' engagement in creative processes. *J. Manage*. 30, 453–470. doi: 10.1016/j.jm.2003.07.001

Gray, M. L., and Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley From Building a New Global Underclass*. New York, NY: Eamon Dolan Books.

Haas, M., and Mortensen, M. (2016). The secrets of great teamwork. *Harvard Bus. Rev*. 94, 70–76.

Hack the Crisis Team. (2021). *Hack The Crisis Join the Brightest Minds to Tackle COVID-19*. Available online at: https://www.hackthecrisis.nl/en/#about (accessed September 17, 2021).

Haeussler, C., and Sauermann, H. (2020). Division of labor in collaborative knowledge production: the role of team size and interdisciplinarity. *Res. Policy* 49,103987. doi: 10.1016/j.respol.2020.103987

Hasteer, N., Bansal, A., and Murthy, B. (2015). An agent based simulation study of association amongst contestants in crowdsourcing software development through preferential attachment. *J. Eng. Appl. Sci*. 10, 2509–2517.

Haun, D., and Over, H. (2015). "Like me: a homophily-based account of human culture," in *Epistemological Dimensions of Evolutionary Psychology*, ed T. Breyer (New York, NY: Springer), 117–130. doi: 10.1007/978-1-4939-1387-9_6

Highsmith, J. (2009). *Agile Project Management: Creating Innovative Products*. Boston: Pearson education.

Jackson, S. E. (1983). Participation in decision making as a strategy for reducing job-related strain. *J. Appl. Psychol*. 68,3. doi: 10.1037/0021-9010.68.1.3

Jarrahi, M. H., Sutherland, W., Nelson, S. B., and Sawyer, S. (2020). Platformic management, boundary resources for gig work, and worker autonomy. *Comput. Support. Cooper. Work* 29, 153–189. doi: 10.1007/s10606-019-09368-7

Jiang, J., An, B., Jiang, Y., Zhang, C., Bu, Z., and Cao, J. (2019). Group-oriented task allocation for crowdsourcing in social networks. *IEEE Trans. Syst. Man Cybern*. 51, 4417–4432. doi: 10.1109/TSMC.2019.2933327

Juárez, J., Santos, C., and Brizuela, C. A. (2021). A comprehensive review and a taxonomy proposal of team formation problems. *ACM Comput. Surveys* 54, 1–33. doi: 10.1145/3465399

Kenna, R., and Berche, B. (2012). Managing research quality: critical mass and optimal academic research group size. *IMA J. Manage. Math*. 23, 195–207. doi: 10.1093/imaman/dpr021

Khan, V.-J., Papangelis, K., Lykourentzou, I., and Markopoulos, P. (2019). *Macrotask Crowdsourcing*. Cham: Springer International Publishing. doi: 10.1007/978-3-030-12334-5

Kittur, A., and Kraut, R. E. (2008). "Harnessing the wisdom of crowds in wikipedia: quality through coordination," in *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA), 37–46. doi: 10.1145/1460563.1460572

Lakhani, K. R., Fayard, A.-L., Levina, N., and Pokrywa, S. H. (2012). OpenIDEO. *Harvard Bus. Sch. Technol. Operat. Mgt. Unit Case*. 612–666.

Lakhani, K. R., and Lonstein, E. (2008). *InnoCentive.com (A)*. Boston, MA: Harvard Business School case. 608, 170.

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett*. 87, 198701. doi: 10.1103/PhysRevLett.87.198701

Lawler, E., and Worley, C. (2009). Designing organizations that are built to change. *Organ. Fut*. 2, 188–202.

Liu, Q., Luo, T., Tang, R., and Bressan, S. (2015). "An efficient and truthful pricing mechanism for team formation in crowdsourcing markets," in *2015 IEEE*

International Conference on Communications (ICC), London: IEEE. 567–572. doi: 10.1109/ICC.2015.7248382

LLP, D. T. T. I. (2020). *Future of Work Accelerated: Learnings From the Covid-19 Pandemic*. Available online at: https://www2.deloitte.com/content/dam/Deloitte/in/Documents/human-capital/in-consulting-accelerated-hc-consulting-noexp.pdf (accessed April 1, 2022).

Lykourentzou, I., Antoniou, A., Naudet, Y., and Dow, S. P. (2016a). "Personality matters: Balancing for personality types leads to better outcomes for crowd teams," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work* & *Social Computing* (San Francisco), 260–273. doi: 10.1145/2818048.2819979

Lykourentzou, I., Liapis, A., Papastathis, C., Papangelis, K., and Vassilakis, C. (2019). "Exploring self-organisation in crowd teams," in *Conference on e-Business, e-Services and e-Society* (Cham; Trondheim: Springer), 164–175. doi: 10.1007/978-3-030-39634-3_15

Lykourentzou, I., Papadaki, K., Vergados, D. J., Polemi, D., and Loumos, V. (2010). Corpwiki: a self-regulating wiki to promote corporate collective intelligence through expert peer matching. *Inform. Sci*. 180, 18–38. doi: 10.1016/j.ins.2009.08.003

Lykourentzou, I., Vinella, F. L., Ahmed, F., Papastathis, C., Papangelis, K., Khan, V.-J., et al. (2021). Self-organizing teams in online work settings. *arXiv[Preprint]. arXiv:2102.07421*. doi: 10.48550/arXiv.2102.07421

Lykourentzou, I., Wang, S., Kraut, R. E., and Dow, S. P. (2016b). "Team dating: a self-organized team formation strategy for collaborative crowdsourcing," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA), 1243–1249. doi: 10.1145/2851581.2892421

Manyika, J., Lund, S., Bughin, J., Robinson, K., Mischke, J., and Mahajan, D. (2016). *Independent-Work-Choice-Necessity-and-the-Gig-Economy*. Technical Report, McKinsey Global Institute.

Marston, W. M. (2013). *Emotions of Normal People*. Oxon: Routledge. doi: 10.4324/9781315010366

Martius, G., and Herrmann, J. M. (2012). Variants of guided self-organization for robot control. *Theory Biosci*. 131, 129–137. doi: 10.1007/s12064-011-0141-0

Marwell, G., Oliver, P. E., and Prahl, R. (1988). Social networks and collective action: a theory of the critical mass. III. *Am. J. Sociol*. 94, 502–534. doi: 10.1086/229028

Marzo Serugendo, G. D., Foukia, N., Hassas, S., Karageorgos, A., Mostéfaoui, S. K., Rana, O. F., et al. (2003). "Self-organisation: paradigms and applications," in *International Workshop on Engineering Self-Organising Applications* (Berlin; Heidelberg: Springer), 1–19. doi: 10.1007/978-3-540-24701-2_1

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Annu. Rev. Social*. 27, 415–444. doi: 10.1146/annurev.soc.27.1.415

Moe, N. B., and Dingsøyr, T. (2008). "Scrum and team effectiveness: theory and practice," in *International Conference on Agile Processes and Extreme Programming in Software Engineering* (Berlin; Heidelberg: Springer), 11–20. Springer. doi: 10.1007/978-3-540-68255-4_2

Monsef, S., Javanmard, S. H., Amini-Rarani, M., Yarmohammadian, M. H., Yazdi, Y., and Haghshenas, A. (2021). Idea generation through Hackathon event in emergencies and disasters, with emphasis on managing flash flood disaster. *Disast. Med. Publ. Health Prepared*. 15, 1–5. doi: 10.1017/dmp.2021.30

Moreland, R. L. (2010). Are dyads really groups? *Small Group Res*. 41, 251–267. doi: 10.1177/1046496409358618

Nolte, A., Pe-Than, E. P. P., Filippova, A., Bird, C., Scallen, S., and Herbsleb, J. D. (2018). You hacked and now what? -exploring outcomes of a corporate Hackathon. *Proc. ACM Hum. Comput. Interact*. 2, 1–23. doi: 10.1145/3274398

Nurzaman, S. G., Yu, X., Kim, Y., and Iida, F. (2014). Guided self-organization in a dynamic embodied system based on attractor selection mechanism. *Entropy* 16, 2592–2610. doi: 10.3390/e16052592

Ortu, M., Destefanis, G., Counsell, S., Swift, S., Tonelli, R., and Marchesi, M. (2017). How diverse is your team? Investigating gender and nationality diversity in Github teams. *J. Softw. Eng. Res. Dev*. 5, 1–18. doi: 10.1186/s40411-017-0044-y

Popescu, G. H., Petrescu, I. E., and Sabie, O. M. (2018). Algorithmic labor in the platform economy: digital infrastructures, job quality, and workplace surveillance. *Econ. Manage. Financ. Mark*. 13, 74–79. doi: 10.22381/EMFM13320184

Prokopenko, M. (2009). *Guided Self-Organization*. Taylor & Francis. doi: 10.2976/1.3233933

Prolific Team. (2021). *Prolific Demographics of Participant Pool*. Available online at: https://researcher-help.prolific.co/hc/en-gb/articles/360009391633-Exporting-Prolific-Demographic-Data (accessed September 21, 2021).

Rahman, H., Roy, S. B., Thirumuruganathan, S., Amer-Yahia, S., and Das, G. (2019). Optimized group formation for solving collaborative tasks. *VLDB J*. 28, 1–23. doi: 10.1007/s00778-018-0516-7

Ramadi, K. B., and Nguyen, F. T. (2021). Rapid crowdsourced innovation for covid-19 response and economic growth. *NPJ Digit. Med*. 4, 1–5. doi: 10.1038/s41746-021-00397-5

Rasmussen, T. H., and Jeppesen, H. J. (2006). Teamwork and associated psychological factors: a review. *Work Stress* 20, 105–128. doi: 10.1080/02678370600920262

Retelny, D., Bernstein, M. S., and Valentine, M. A. (2017). No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proc. ACM Hum. Comput. Interact*. 1, 1–23. doi: 10.1145/3134724

Retelny, D., Robaszkiewicz, S., To, A., Lasecki, W. S., Patel, J., Rahmati, N., et al. (2014). "Expert crowdsourcing with flash teams," in *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, HI), 75–85. doi: 10.1145/2642918.2647409

Rokicki, M., Zerr, S., and Siersdorfer, S. (2015). "Groupsourcing: team competition designs for crowdsourcing," in *Proceedings of the 24th International Conference on World Wide Web* (Florence), 906–915. doi: 10.1145/2736277.2741097

Roy, S., Balamurugan, C., and Gujar, S. (2013). "Sustainable employment in India by crowdsourcing enterprise tasks," in *Proceedings of the 3rd ACM Symposium on Computing for Development* (Bangalore), 1–2. doi: 10.1145/2442882.2442904

Salehi, N., and Bernstein, M. S. (2018). Hive: collective design through network rotation. *Proc. ACM Hum. Comput. Interact*. 2, 1–26. doi: 10.1145/3274420

Schriner, A., and Oerther, D. (2014). No really (crowd) work is the silver bullet. *Proc. Eng*. 78, 224–228. doi: 10.1016/j.proeng.2014.07.060

Silberman, M. S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., and Zaldivar, A. (2018). Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 39–41. doi: 10.1145/3180492

Smith, R., and Leberstein, S. (2015). *Rights on Demand. Ensuring Workplace Standards and Worker Security In the On-Demand Economy*. National Employment Law Project, Washington, DC.

Taha, H. A. (2013). *Operations-Research-An-Introduction-10th-Ed*. Harlow.

Temiz, S. (2021). Open innovation *via* crowdsourcing: a digital only Hackathon case study from Sweden. *J. Open Innov*. 7:39. doi: 10.3390/joitmc7010039

Valentine, M. A., Retelny, D., To, A., Rahmati, N., Doshi, T., and Bernstein, M. S. (2017). "Flash organizations: crowdsourcing complex work by structuring crowds as organizations," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY), 3523–3537. doi: 10.1145/3025453.3025811

Wang, R. (2020). Marginality and team building in collaborative crowdsourcing. *Online Inform. Rev*. 44, 827–846. doi: 10.1108/OIR-09-2018-0269

Wang, S., Yeoh, W., Ren, J., and Lee, A. (2021). Learnings and implications of virtual Hackathon. *J. Comput. Inform. Syst*. 62, 1–13. doi: 10.1080/08874417.2020.1864679

Whiting, M. E., Gamage, D., Goyal, S., Gilbee, A., Majeti, D., Richmond-Fuller, A., et al. (2017). "Designing a constitution for a self-governing crowdsourcing marketplace," in *Collective Intelligence Conference* (Brooklyn, NY), 15–16.

Whitson, J. R., Simon, B., and Parker, F. (2021). The missing producer: rethinking indie cultural production in terms of entrepreneurship, relational labour, and sustainability. *Eur. J. Cult. Stud*. 24, 606–627. doi: 10.1177/1367549418810082

Wood, A. J., Graham, M., Lehdonvirta, V., and Hjorth, I. (2019). Networked but commodified: the (dis) embeddedness of digital labour in the gig economy. *Sociology* 53, 931–950. doi: 10.1177/0038038519828906

Wu, G., Chen, Z., Liu, J., Han, D., and Qiao, B. (2021). Task assignment for social-oriented crowdsourcing. *Front. Comput. Sci*. 15,8. doi: 10.1007/s11704-019-9119-8

Yates, F. E. (2012). *Self-Organizing Systems: The Emergence of Order*. Denver, Colorado: Springer Science & Business Media.

Yin, M., Suri, S., and Gray, M. L. (2018). "Running out of time: the impact and value of flexibility in on-demand crowdwork," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal), 1–11. doi: 10.1145/3173574.3174004

Yin, X., Wang, H., Wang, W., and Zhu, K. (2020). Task recommendation in crowdsourcing systems: a bibliometric analysis. *Technol. Soc*. 63:101337. doi: 10.1016/j.techsoc.2020.101337

Yu, D., Zhou, Z., and Wang, Y. (2019). Crowdsourcing software task assignment method for collaborative development. *IEEE Access* 7, 35743–35754. doi: 10.1109/ACCESS.2019.2905054

Zhou, S., Valentine, M., and Bernstein, M. S. (2018). "In search of the dream team: temporally constrained multi-armed bandits for identifying effective team structures," in *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems* (Montreal), 1–13. doi: 10.1145/3173574.3173682

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/frai.2022.828733

An Analysis of Music Perception Skills on Crowdsourcing Platforms

Ioannis Petros Samiotis[1]*, *Sihang Qiu*[1,2], *Christoph Lofi*[1], *Jie Yang*[1], *Ujwal Gadiraju*[1] and *Alessandro Bozzon*[1]

[1] *Department of Software Technology, Delft University of Technology, Delft, Netherlands,* [2] *Hunan Institute of Advanced Technology, Changsha, China*

Music content annotation campaigns are common on paid crowdsourcing platforms. Crowd workers are expected to annotate complex music artifacts, a task often demanding specialized skills and expertise, thus selecting the right participants is crucial for campaign success. However, there is a general lack of deeper understanding of the distribution of musical skills, and especially auditory perception skills, in the worker population. To address this knowledge gap, we conducted a user study ($N = 200$) on Prolific and Amazon Mechanical Turk. We asked crowd workers to indicate their musical sophistication through a questionnaire and assessed their music perception skills through an audio-based skill test. The goal of this work is to better understand the extent to which crowd workers possess higher perceptions skills, beyond their own musical education level and self reported abilities. Our study shows that untrained crowd workers can possess high perception skills on the music elements of *melody*, *tuning*, *accent*, and *tempo*; skills that can be useful in a plethora of annotation tasks in the music domain.

Keywords: human computation, music annotation, perceptual skills, music sophistication, knowledge crowdsourcing

OPEN ACCESS

Edited by:
Rashid Mehmood,
King Abdulaziz University, Saudi Arabia

Reviewed by:
Rocco Zaccagnino,
University of Salerno, Italy
F. Amilcar Cardoso,
University of Coimbra, Portugal

****Correspondence:***
Ioannis Petros Samiotis
i.p.samiotis@tudelft.nl

Specialty section:
This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

Received: *03 December 2021*
Accepted: *09 May 2022*
Published: *14 June 2022*

Citation:
Samiotis IP, Qiu S, Lofi C, Yang J, Gadiraju U and Bozzon A (2022) An Analysis of Music Perception Skills on Crowdsourcing Platforms.
Front. Artif. Intell. 5:828733.
doi: 10.3389/frai.2022.828733

1. INTRODUCTION

Several studies have shown the ability of crowd workers to successfully contribute to the analysis and annotation of multimedia content, both based on simple perceptual skill, e.g., for image analysis (Sorokin and Forsyth, 2008) and domain-specific knowledge, (Oosterman et al., 2015). Musical content is no exception, and research has shown that the general crowd can be successfully involved in the annotation (Samiotis et al., 2020) and evaluation (Urbano et al., 2010) processes of music-related data and methods. Plenty of music annotation tasks, (Lee, 2010; Mandel et al., 2010; Speck et al., 2011; Lee and Hu, 2012; Lee et al., 2012) can be routinely found on microtask crowdsourcing platforms, mostly focused on descriptive (Law et al., 2007) and emotional (Lee, 2010) tagging.

Music, as a form of art, often requires a multifaceted set of skills to perform and certain expertise to analyse its artifacts. There are cases that require advanced music perceptual skills (such as the ability to perceive changes in melody) and music-specific knowledge. However, both in literature and in practice, it is rare to encounter such crowdsourcing tasks. Consider, for example, annotation tasks targeting classical music, e.g., music transcription, performance evaluation, or performance annotation. Classical music is a genre featuring artworks with high musical complexity; it is no surprise that corresponding analysis and annotation tasks are often exclusively performed by musical experts and scholars. This unfortunately hampers current efforts to digitize and open up classical music archives, as scholars and experts are expensive and not easily available. Here, the ability to utilize microtask crowdsourcing as an annotation and analysis approach could bring obvious advantages. But how likely it is to find advanced music-related perceptual skills on crowdsourcing platforms? With the goal of answering this broad research question, in this paper we scope our investigation on the following two aspects:

- **[RQ1]** How are different music perception skills and self-reported music-related knowledge distributed among crowd workers of different platforms?
- **[RQ2]** How are music perception skills associated to domain and demographic attributes?

Studies on human cognition and psychology have shown that people can possess innate music perception skills without previous formal training (Ullén et al., 2014; Mankel and Bidelman, 2018). However, the majority of those studies have been conducted in labs, under controlled conditions and with limited amounts of participants.

In our work, we set out to measure the music sophistication and perception skills of crowd workers operating on the Prolific[1] and Amazon Mechanical Turk[2] crowdsourcing platforms. We chose to conduct our study on these two different platforms, in order to diversify our participant pool and identify potential differences between them. In its present form, this study expands the preliminary study as presented in Samiotis et al. (2021), by diversifying the participant pool and complementing the analysis with additional methods.

We designed a rigorous study that employs validated tools to measure the musical sophistication of the users and quantify their music perception skills: the Goldsmith's Music Sophistication Index (GMSI) questionnaire (Müllensiefen et al., 2014) and the Profile of Music Perception Skills (PROMS) active skill test (Law and Zentner, 2012), respectively (and more specifically its shorten version: Mini-PROMS). These tools allow for a general overview of musical ability characteristics, but also a more detailed understanding through their subcategories (e.g., musical training and melody perception skills). By juxtaposing passive methods of assessment (questionnaire) with the active evaluation of auditory skills, we aim to gather a better understanding of workers' actual skills on musical aspects, beyond their subjective self-assessment. With GMSI, we are able to evaluate a person's ability to engage with music through a series of questions focusing on different musical aspects. PROMS on the other hand, allows for a more objective way to measure a person's auditory music perception skills (e.g., melody, tuning, accent, and tempo perception) through a series of audio comparison tests. To the best of our knowledge, this is the first attempt to use PROMS in an online crowdsourcing environment and the measured perception skills can offer valuable insights to the auditory capabilities of the crowd.

Our findings indicate that pre-existing musical training is not common among crowd workers, and that music sophistication aspects are not necessarily predictive of actual music perception skills. Instead, we observe that the majority of workers show an affinity with specific sets of skills (e.g., we found a surprising number of *musical sleepers* — workers without formal training but still high music perception skill test results). As a whole, our study paves the way for further work in worker modeling and task assignment, to allow a wider and more refined set of microtask crowdsourcing tasks in the domain of music analysis and annotation.

2. RELATED WORK

There is a long history of studies on perception and processing of music by humans; from the analysis of the socio-cultural variables influencing a person's musicality amplitude (Hannon and Trainor, 2007), to the study of musicality from a genetics' base (Gingras et al., 2015). In all cases, inherent music processing capabilities have been found in people and they seem to be connected with basic cognitive and neural processes of language since early stages of development (Liberman and Mattingly, 1985; Koelsch et al., 2009). Even people with *amusia*, a rare phenomenon where a person can't distinguish tonal differences between sounds (Peretz and Hyde, 2003), they can still process and replicate rhythm correctly (Hyde and Peretz, 2004).

In Müllensiefen et al. (2014), we find a large scale study on musical sophistication through the use of the GMSI survey, on a unique sample of 147,663 people. GMSI is particularly calibrated to identify musicality in adults with varying levels of formal training. It is targeted toward the general public, and can prove less effective to distinguish fine differences between highly trained individuals. Musical sophistication in the context of that study, and ours, encompasses musical behaviors and practices that go beyond formal training on music theory and instrument performance. Their findings show that musical sophistication, melody memory and musical beat perception are related. The survey has been translated and replicated successfully (on smaller samples) in French (Degrave and Dedonder, 2019), Portuguese (Lima et al., 2020), Mandarine (Lin et al., 2021), and German (Schaal et al., 2014).

Our study draws connections to those findings and aims to shed light into the musical capabilities of people on crowdsourcing platforms. The demographics and conditions of the studies presented so far cannot be easily compared to those of online markets. Users on those platforms are participating in such studies through monetary incentives, and the conditions (equipment, location, potential distractions, etc.) under which they perform the tasks cannot be controlled as in a lab environment, as indicated in Totterdell and Niven (2014), Gadiraju et al. (2017), and Zhuang and Gadiraju (2019).

Currently, crowdsourced music annotation is primarily utilized for descriptive (Law et al., 2007) and emotional (Lee, 2010) tagging. Large-scale music data creation and annotation projects such as Last.fm[3] and Musicbrainz[4], are largely depended on human annotation, but from users of their respective online social platforms. A survey on the applicability of music perception experiments on Amazon Mechanical Turk (Oh and Wang, 2012), showed that online crowdsourcing platforms have been underused in the music domain and the status has not changed radically since then. Through our study, we want to

[1] https://www.prolific.co
[2] https://https://www.mturk.com
[3] https://www.last.fm
[4] https://musicbrainz.org

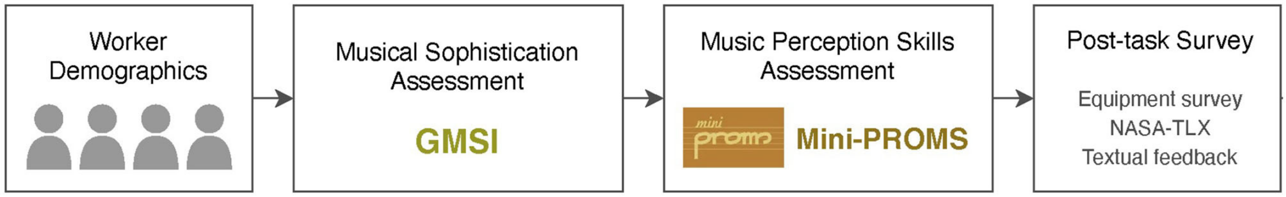


FIGURE 1 | The four steps in the music perception skills study.

examine the capabilities of the crowd on processing music audio and showcase their capabilities, in an attempt to encourage further research and utilization of crowdsourcing in the music domain. Although our focus on audio perception separates our work from visual-based studies on music perception, it is meaningful to mention that visualization techniques for music tasks have proven effective for certain use cases such as music plagiarism detection in De Prisco et al. (2016) and De Prisco et al. (2017) but also harmonic structure perception in music, in Malandrino et al. (2015).

3. EXPERIMENTAL DESIGN

The main focus of this study is to offer insights into the musical characteristics and perception skills of workers operating on crowdsourcing platforms. We therefore designed our experiment to capture these attributes through methods that can be used online, and that do not require pre-existing musical knowledge. We used two methods: 1) the *GMSI* questionnaire to evaluate the *musical sophistication* (musical training, active engagement and other related musical characteristics) (Müllensiefen et al., 2014) of workers and 2) the *Mini-PROMS* test battery to evaluate their auditory music perception skills. We then compare the obtained results, paying specific attention to the overlapping aspects of musical sophistication and music perception skills. With this experiment, we are also interested in identifying "*musical sleepers*" and "*sleeping musicians*", a notion originally presented in Law and Zentner (2012). A musical sleeper is a person with little to no musical training but with high performance in the perception test, while a sleeping musician indicates the opposite.

3.1. Procedure

After a preliminary step where workers are asked basic demographic information (age, education, and occupation), the study is composed of four consecutive steps (**Figure 1**), each devoted to collecting information about specific attributes corresponding to the crowd workers: (1) Musical Sophistication Assessment (*GSMI*), (2) Active Music Perception Skill Assessment (*Mini-PROMS*) and (3) Post-task Survey collecting information on workers audio-related conditions, and perceived cognitive load.

3.2. Questionnaires and Measures

3.2.1. Capturing Musical Sophistication of Workers

Musical behaviors of people such as listening to music, practicing an instrument, singing or investing on vinyl collections, all show the affinity of a person toward music. The degree to which a person is engaged to music through these behaviors, constitutes the musical sophistication. Musical sophistication can be measured as a psychometric construct through the *GMSI* questionnaire, which collects self-reported musicality through emotional responses, engagement with music, formal training, singing capabilities and self-assessed perception skills. It is an instrument specifically designed to capture the sophistication of musical behaviors, in contrast to other questionnaires such as Musical Engagement Questionnaire (MEQ) (Werner et al., 2006), which measures the spectrum of psychological facets of musical experiences. More specifically, the musical sophistication of people based on Müllensiefen et al. (2014), is organized into the following five facets:

- Active Engagement: this aspect determines the degree to which a person engages with music, by listening to and allocating their time/budget to it;
- Perceptual Abilities: this aspect assesses the skill of perceiving (mainly auditory) elements of music. This is an important subscale in our study, since the self-assessed perceptual skills of the workers in GMSI can be directly compared to those we actively measure in Mini-PROMS;
- Musical Training: this aspect reports the years of training on aspects of music (e.g., theory, performing an instrument), which can indicate the formal expertise that a person has in the domain;
- Emotions: this aspect determines the emotional impact of music on that person;
- Singing Abilities: this aspect evaluates the ability to follow along melodies and tempo (beat) of songs.

GMSI offers additional questions outside the subscales, which capture specific properties of the participant: 1) "Best Instrument", which represents which instrument the user knows to play the best, 2) "Start Age", which age the participant starting learning an instrument and 3) "Absolute Pitch", which indicates if the person can understand correctly the exact notes of a sound frequency. Absolute pitch is a very rare trait that develops during the early stages of auditory processing (Burkhard et al., 2019) but can deteriorate through the years (Baharloo et al., 1998). As such, a person with perfect pitch perception, could have an advantage on a melody perception test, thus we included it with the rest of the subscales.

The original GMSI questionnaire contains 38 main items and 3 special questions, and considering the rest of the study's parts, we chose to reduce its size while keeping its psychometric reliability. For that purpose, we consulted the GSMI online "configurator"[5] which allows to select the number of items per subscales and estimates the reliability of the resulting questionnaire based on the questions it selects. We reduced the size of the questionnaire to 34 questions, and preserved the special question about "Absolute Pitch", resulting in 35 questions in total.

In the GMSI questionnaire each question from the subscales uses the seven-point Likert scale (Joshi et al., 2015) for the user's responses, with most questions having "Completely Agree", "Strongly Agree", "Agree", "Neither Agree Nor Disagree", "Disagree", "Strongly Disagree" and "Completely Disagree" as options. Few questions offer numerical options for topics (e.g., indicating the time spent actively listening to music, or practicing an instrument). The workers are not aware of the subscale each question belongs to. The index of each subscale of GMSI is calculated with the aggregated results of the relevant questions. The overall index of "General Music Sophistication" is calculated based on 18 questions out of the total 34 items of the subscales; these 18 questions are predefined by the designers of the questionnaire; the question about "Absolute Pitch" does not contribute to the total index.

Using the GMSI questionnaire is close to the typical methods used to assess the knowledge background of annotators in other domains. Especially the questions of "Musical Training" follow standard patterns to assess the formal training of a person in a domain, thus a certain objectivity can be expected (assuming good faith from the workers). However, the rest of the categories are based purely on subjective indicators and self-reported competence, which can potentially misrepresent the true music behaviors and capabilities of a worker. For this reason, it is necessary to understand the best practices that could reliably predict a worker's performance to a music annotation task. To that end, we compare the workers' input in such questionnaires, and specifically on GMSI, to the music perceptual skills they might possess, which we measure through an audio-based, music perception skill-test.

3.2.2. Measuring Music Perception Skills of Workers

The music perception skill test is based on the well-establish *Profile of Music Perception Skills* (PROMS) test (Law and Zentner, 2012). Its original version is quite extensive and its completion can take more than an hour, as it covers several music cognition aspects like Loudness, Standard rhythm, Rhythm-to-melody, Timbre, Pitch and more. Considering the possibly low familiarity of crowd workers with these tasks and its inherent difficulty, we opted for a shorter version, the *Mini-PROMS* (Zentner and Strauss, 2017), which has also been adopted and validate in the context of online, uncontrolled studies.

Mini-PROMS is a much shorter battery of tests (15 min completion time), which still covers the "Sequential" and "Sensory" subtests. It can measure a person's music perception skills, by testing their capability to indicate differences on the following musical features:

- Melody: A sequence of notes, with varying density and atonality
- Accent: The emphasis of certain notes in a rhythmic pattern
- Tuning: The certain frequency of notes, when played in a chord
- Tempo: The speed of a rhythmic pattern.

The musical aspects selected in this test are argued to well represent the overall music perception skills of a person, only in a more concise way. This version retains test–retest reliability and internal consistency values close to the original PROMS test (Law and Zentner, 2012), validating it for our research purposes. Note that, although reduced in size, these four skills are required to enable a broad range of music-related research, such as beat tracking, tonal description, performance assessment and more.

For each of the 4 musical aspects workers receive a brief explanation and an example case to familiarize the user with the test. Each test after the introduction presents a reference audio sample twice and a comparison sample once. The two audio samples can differ based on the musical aspect tested and the worker is asked if the samples are indeed same or differ. The authors of PROMS have put particular effort on distinguishing the musical aspects from each other, to make the skill evaluation as close as possible to the musical aspect tested. Finally, to minimize cognitive biases due to enculturation (Demorest et al., 2008) the audio samples have been created using less popular instrument sounds, such as harpsichord and "rim shots". Meanwhile, the structure of audio samples and the aspect separation allow for a more precise measurement of a person's perception skill.

The categories of "Melody" and "Accent" have 10 comparisons each, while "Tuning" and "Tempo" have 8. After the user has listened to the audio samples, they are asked to select between "Definitely Same", "Probably Same", "I don't know", "Probably Different", and "Definitely Different". The participant is then rewarded with 1 point for the high-confidence correct answer, while the low-confidence one rewards 0.5 point. The subscale scores are calculated through a sum of all items within the scale and divided by 2. The total score is an aggregated result of all subscale scores. During the test, the user is fully aware of the subscale they are tested for, but the name of "Tempo" is presented as "Speed" (original creators' design choice).

3.2.3. Self-Assessment on Music Perception Skills

Self-assessment can often misrepresent an individual's real abilities (Kruger and Dunning, 1999). For that reason, we employed a survey to study this effect its manifestation with music-related skills. After Mini-PROMS test, the worker has to input how many of the comparisons per subscale they believe they correctly completed—this information is not known to them after executing the Mini-PROMS test. Therefore, they are presented with 4 questions, where they have to indicate between 0 and the total number of tests per subscale (10 for "Melody"/"Accent" and 8 for "Tuning"/"Tempo"). Finally, the results of this survey are compared to the score of workers on the "Perceptual Abilities" subscale of GMSI, which also relies

[5] https://shiny.gold-msi.org/gmsiconfigurator/

on self-assessment. We expect workers to re-evaluate their own skills, once exposed to the perception skill test.

3.2.4. Post-task Survey

As a final step of the task, the worker is presented with three post-task surveys: (1) a survey on the audio equipment and the noise levels around them, (2) a survey on the cognitive load they perceived and (3) an open-ended feedback form.

The audio equipment survey consisted of four main questions, to retrieve the type of equipment, its condition and the levels of noise around them during the audio tests. Insights on these can help us understand the to what extent the equipment/noise conditions affected Mini-Proms test, which is audio-based. More specifically, we asked the following questions:

1. What audio equipment were you using during the music skill test?
2. What was the condition of your audio equipment?
3. Does your audio equipment have any impairment?
4. How noisy was the environment around you?

The options regarding the audio equipment were: "Headphones", "Earphones", "Laptop Speakers", and "Dedicated Speakers". For the condition questions (2) and (3), we used the unipolar discrete five-grade scales introduced in ITU-R BS (2003), to subjectively assess the sound quality of the participants' equipment. Finally, for question (4) on noise levels, we used the loudness subjective rating scale, introduced in Beach et al. (2012).

In the second part of post-task survey, the workers had to indicate their cognitive task load, through the NASA's Task Load IndeX (NASA-TLX) survey[6]. The survey contains six dimensions—Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Workers use a slider (ranging from 0 to 20, and later scaled to 0 to 100) to report their feelings for each of the six dimensions. A low TLX score represents the music skill test is not mentally, physically, and temporally demanding, and it also indicates less effort, and less frustration perceived by the worker, while completing the entire study.

Finally, we introduced an free-form textual feedback page, where users were encouraged to leave any comments, remarks, or suggestions for our study.

3.3. Worker Interface

The worker interfaces of our study is using VueJS[7], a JavaScript framework. The first page of our study contained general instructions for the study alongside estimated completion times for each part of it. Each page thereafter, contained an interface for each of the steps in our study, as seen in **Figure 1**.

To assist navigation through the GMSI questionnaire, we implemented the questionnaire interface to show one question at a time. We added a small drifting animation to show the next question, when they select their answer in the previous one. We also added a "back" button, in case they wanted to return to a previous question and alter their answer. They could track their progress through the questionnaire from an indication of the number of the question and the total number of questions (see **Figure 2**).

While we retrieved the questions for GMSI and implemented them in our study's codebase, for PROMS we wanted to use the exact conditions and audio-samples as in Zentner and Strauss (2017). To replicate their test faithfully, the creators of PROMS (Law and Zentner, 2012) kindly gave us access to their Mini-PROMS interfaces (example interface in **Figure 2**). Mini-PROMS is implemented on LimeSurvey[8] and users were redirected to it after the completion of GMSI.

After the GMSI questionnaire, workers were introduced to the page seen in **Figure 2**. There, they had to copy their Participant ID (retrieved programmatically from the crowdsourcing platforms) and use it in the Mini-PROMS interface later, so we could link their test performance (stored in LimeSurvey), with their entries in our database. At the end of Mini-PROMS, the users were redirected back to our study through a provided URL.

In the final stage of our study, the participants were greeted and provided a "completion code", which they could submit on back on their respective platform, to complete the task.

3.4. Participants, Quality Control, and Rewards

On Prolific, we recruited 100 crowd workers to complete our study. We applied a participant selection rule for "Language Fluency": English, as all of our interfaces were implemented in English. Only crowd workers whose overall approval rates were higher than 90% could preview and perform our study. On Amazong Mechanical Turk, we recruited 100 crowd workers as well, where we set their approval rate to "greater than 90%".

To assess the quality of the user input, we included attention check questions on the GMSI and NASA-TLX interfaces of the study. More specifically, we included three attention check questions in GMSI, asking the participants to select a specific item in the same seven-point Likert scale. In the NASA-TLX survey, we included a question asking the users to select a specific value out of the 21 available in the scale of the survey.

We set the reward on Prolific and Amazon Mechanical Turk for completing our study to 3.75 GBP (5.2 USD). Upon the completion of our study on both platforms, workers immediately received the reward. The average execution time was 32.5 min, resulting in the hourly wage of 7.5 GBP (10.3 USD), rated as a "good" pay by the platforms.

4. RESULTS

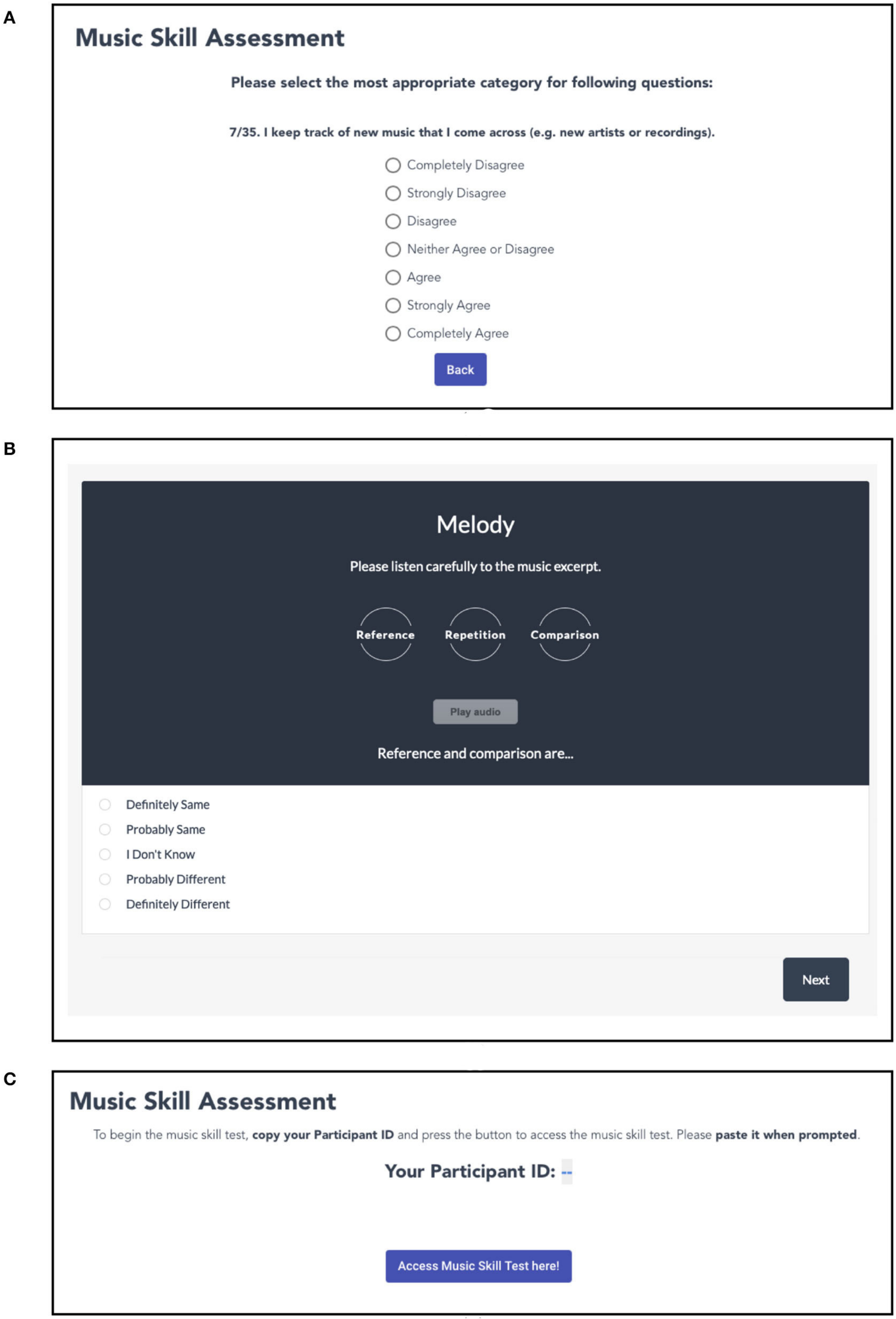
While investigating the data we gathered in our study, we followed similar analysis steps for both platforms. The data were first cleaned up based on our attention check questions and we only kept demographic data that we had actively asked the participants (dropped platform-based demographics).

We proceeded with identifying the distribution characteristics of each variable from the different parts of our study (GMSI

[6] https://humansystems.arc.nasa.gov/groups/tlx/

[7] https://vuejs.org

[8] https://www.limesurvey.org



Music Skill Assessment

Please select the most appropriate category for following questions:

7/35. I keep track of new music that I come across (e.g. new artists or recordings).

○ Completely Disagree
○ Strongly Disagree
○ Disagree
○ Neither Agree or Disagree
○ Agree
○ Strongly Agree
○ Completely Agree

Melody

Please listen carefully to the music excerpt.

Reference and comparison are...

○ Definitely Same
○ Probably Same
○ I Don't Know
○ Probably Different
○ Definitely Different

Music Skill Assessment

To begin the music skill test, **copy your Participant ID** and press the button to access the music skill test. Please **paste it when prompted**.

Your Participant ID: --

FIGURE 2 | Interfaces of the study (**A**, GMSI questionnaire, **B**, Mini-PROMS, and **C**, Participant ID prompt).

TABLE 1 | Prolific participant demographics.

	Variables	Statistics
Age (years)	Range	18–65
	Majority	18–25 (70.11%)
Occupation	Full-time	30
	Part-time	11
	Unemployed	44
	Voluntary work	2
Education	Associate degree	3
	Bachelor's degree	35
	Doctorate degree	1
	High school/HED	16
	Master's degree	12
	Professional degree	1
	Some college, no diploma	13
	Some high school, no diploma	2
	Technical/trade/vocational training	4

TABLE 2 | GMSI range, median, mean, and standard deviation.

	Range	Median	Mean	Standard deviation (1σ)
Active engagement	19–45	31	30.91	5.45
Perceptual abilities	16–45	34	33.62	6.65
Musical training	7–45	17	18.52	9.61
Singing abilities	9–41	28	27.41	6.03
Emotions	18–42	33	33.24	4.28
General music sophistication	40–101	69	69.76	14.20

and Mini-PROMS subcategories, NASA-TLX and equipment questions). Combined with the intercorrelations per study part, we gained important insights on the attributes of each variable and their relations. These results are compared to those of the original GMSI and Mini-PROMS studies, to assess the differences between the different participants' pools. Finally, we run a Multiple Linear Regression, to assess which factors seem to be the best predictors for the music perception skills of a crowd worker (e.g., musical training, equipment quality etc.).

4.1. Prolific

Of the 100 workers recruited from Prolific, 8 of them failed at least one attention check question(s); 5 of them provided invalid/none inputs. After excluding these 13 invalid submissions, we have 87 valid submissions from 87 unique workers.

4.1.1. Worker Demographics

Table 1 summarizes workers' demographic information. Of the 87 crowd workers who provided valid submissions, 36 were female (41.38%), while 51 were male (58.62%). Age of participants ranged between 18 and 58 and the majority of them were younger than 35 (87.36%). The majority of the workers (51%) were reported to be unemployed, while from those employed, 73.17% had a full-time job. Most workers had enrolled for or acquired a degree (78.16%), with 51.47% of them pointing to Bachelor's degree. In total, we employed workers from 15 countries, with most workers (77%) currently residing in Portugal (25), United Kingdom (16), Poland (13), and South Africa (13).

4.1.2. Results on Worker Music Sophistication

Table 2 summarizes the results of the GMSI questionnaire on our workers. We contrast our results to results of the original GMSI study (Müllensiefen et al., 2014), which covered a large population sample of participants $n = 147,663$ that voluntary completed the questionnaire, on BBC's *How Musical Are You?* online test. Participants were mainly UK residents (66.9%) and, in general, from English-speaking countries (USA: 14.2%, Canada: 2.3%, Australia: 1.1%), with 15.9% having non-white background. The sample contained a large spread on education and occupation demographics, where only 1.8% claimed working in the music domain. To some extent, this study is considered representative for the general population in the UK (but is biased toward higher musicality due to the voluntary nature of that study). As such, we can assume a certain disposition and affinity to music from GMSI's population sample, compared to ours where the incentives where monetary.

In our study, the observed General Music Sophistication ($\mu = 69.76$) positions our workers pool at the bottom 28–29% of the general population distribution found in the GMSI study. We observe a similar effect also with the individual subscales with the exception of "Emotions", for which our workers fare a bit higher (bottom 32–38%).

The result indicates that the self-reported music sophistication of crowd workers is strongly below that of the general population. Most workers had received relatively little formal training in their lifetime. This finding is important for the rest of the analysis, as it indicates *low formal expertise* with music among the crowd workers.

Most workers indicate relatively high perceptual abilities ($\mu = 33.62$, $max = 45$). Here, it is interesting that previous studies (Baharloo et al., 2000) estimate that less than 1% (or 5 people) per 11,000 possess "Absolute Pitch". In our sample though, 9 workers indicated having this characteristic, little more than the 10% of our sample. This could indicate a possible confusion between quasi-absolute pitch which is related to the familiarity of a person with an instrument's tuning and timber (Reymore and Hansen, 2020), or with relative pitch. Relative pitch is trainable through practice and useful to professional musicians, as they can detect changes in pitch through the relations of tones (5 out of 9 workers who indicated "Absolute Pitch" had scored higher than 30 out of 49 in the "Musical Training" category scale, indicating adequate formal musical training).

Table 3 presents the correlations between GMSI subscales. As the scores of each GMSI subscale follow a normal distribution (Shapiro-Wilk test), we applied Pearson's R test to calculate correlation coefficients. We observe that Perceptual Abilities

TABLE 3 | Intercorrelations (Pearson's R) of subscales of GMSI scores.

	Active engagement	Perceptual abilities	Musical training	Emotions	Singing abilities
Active engagement	1.000				
Perceptual abilities	0.262*	1.000			
Musical training	0.224*	0.442*	1.000		
Emotions	0.401*	0.380*	0.178	1.000	
Singing abilities	0.142	0.463*	0.465*	0.125	1.000

Statistical significance (p < 0.05) is marked using an asterisk (*).

TABLE 4 | Mini-PROMS range, median, mean, and standard deviation.

	Range	Median	Mean	Standard deviation (1σ)
Melody	1.5–9	5	4.98	1.59
Tuning	1–7.5	4	4.22	1.62
Accent	0–9.5	5	5.19	1.84
Tempo	1–8	5	5.14	1.59
Mini-PROMS total	6–30	19.5	19.53	4.98

shows positive correlations with most other subscales ($p < 0.05$), especially with Music Training ($R = 0.442$), Emotions ($R = 0.380$), and Singing Abilities ($R = 0.463$). This finding suggests that the listening skill plays the most important role in crowd workers' music sophistication. We also find significant correlations between Active Engagement and Emotions ($R = 0.401$), and between Singing Abilities and Musical Training ($R = 0.465$). The original GMSI study has shown that different subscales are strongly correlated ($R > 0.486$). The difference we observe could be partly explained by the generally lower musical sophistication scores of the crowd workers in our pool.

4.1.3. Results on Objective Music Perception Skills

Mini-PROMS categorizes perception skills as "Basic" if the total obtained score is lower than 18, "Good" if between 18 and 22.5, "Excellent" for values between 23 and 27.5, and "Outstanding" for values over 28 (Zentner and Strauss, 2017). The original Mini-PROMS study covered a total $n = 150$ sample of participants, all recruited from the university of Innsbruck, via email. Most of the participants were students with at least one degree ($n = 134$), aged 27 on average.

We observed (see **Table 4**) an average of "Good" music perception skills for our workers ($\mu = 19.53$, avg. accuracy 54.25%). Forty-eight out of eighty-seven (55.17%) produced reasonably high accuracy in music skill tests (belonging to "Good" and better categories according to Mini-PROMS results). These figures are lower compared to the results of the original study (Zentner and Strauss, 2017) ($\mu = 24.56$, 68.2% avg. accuracy), a fact that we account to the greater representation of *non-musician* in our workers pool (67.82%), compared to the participants of the original Mini-PROMS study (where only 38.67% identified as non-musicians). However, considering the low formal training amongst the surveyed workers, we consider this result an indication of the existence of useful and somewhat abundant auditory music perception skills among untrained workers. Especially, in the top 10% of workers, ranked according to their total Mini-PROMS values, several achieved quite high accuracy, between 73.6 and 83.3%, which would indicate perception skills between "Excellent" and "Outstanding" in Mini-PROMS's scale. In the following section we will analyse in greater detail the relationship between the measured music sophistication and the perception skills.

A similar trend toward lower performance compared to the original Mini-PROMS study can be observed across the other musical aspects: workers correctly identified melody differences with 49.77% avg. accuracy (original study: 64.3%), tuning differences with 52.73% avg. accuracy (original: 68%), accent difference with 51.95% avg. accuracy (original study: 61.5%), and tempo differences with 64.3% avg. accuracy (original study: 81.25%).

The result of the music skill tests is in-line with the result of self-reported music sophistication from GMSI, suggesting that when compared to the populations covered by previous studies, crowd workers generally possess less music perception skills. To deepen the analysis, we calculated the intercorrelation of Mini-PROMS subscales, and made comparison with the original study (Zentner and Strauss, 2017). Since the Mini-PROMS scores across all the subscales follow normal distributions based on the Shapiro-Wilk tests (Hanusz et al., 2016), we carried out Pearson's R tests to get the correlation coefficients and corresponding p-values. We find statistical significance on all the intercorrelations. Especially, we find that workers' music skills related to melody are positively correlated with their accent- and tempo-related skills ($R = 0.551$ and $R = 0.514$, respectively), while accent and tempo also shows a moderate correlation ($R = 0.468$). In comparison with the original study, we do not observe large differences in the R values, while we did with the GMSI results. The results of the intercorrelation analysis suggests that worker melody, accent, and tempo skills are related with each other in our population too. This is a positive result, that suggests (1) the applicability of this testing tool also on this population, and (2) the possibility of developing more compact tests for music perception skills, for workers' screening or task assignment purposes.

When focusing on the top 10% of workers, we observed an accuracy on "Melody" between 75% and 90%, while the top 5% scored higher than 85%. A person with "Absolute Pitch" would be expected to achieve high accuracy on this test. Only one person in the top 10% had indicated "Absolute Pitch", but their accuracy was one of the lowest in the group (75%). This could indicate that the person is more likely to not possess such a characteristic. For the subcategory of "Tuning", the top 10% achieved accuracy between 81.25 and 93.75%, while the top 5% scored higher than 87.5%. On "Accent", the top 10% reached accuracy between 80 and 95%. Finally, on the subcategory of "Tempo" we measured accuracy of 87.5 and 100% in the top 10%, while the top 5% achieved perfect score of 100%.

These results suggest the presence of a substantial fraction of workers possessing higher music perception skills than expected from their training, although differently distributed. For example, workers who perceived well changes in "Melody", didn't perform equally well on the other categories. This could indicate that music perception skills do not necessarily "carry over" from one music feature to the other; other workers will be good in perceiving changes in tempo, while others on tuning. This encourages the use of the appropriate set of tests, to identify potentially high performing annotators. Thus, if we take as example beat tracking annotation tasks, it would be more beneficial to focus on testing the rhythm-related perception skills, as the other categories have lower chance to capture the appropriate workers for the task.

4.1.4. Post-task Survey: Equipment and Cognitive Workload

The majority of the workers reported that, during the test, they used headphones (52.87%) (which is very good for musical tasks), earphones (29.54%), and laptop speakers (16.09%) (which are not optimal). All workers reported the quality of their equipment as "Fair" or better quality (55.17% selected "Excellent" and 34.48% "Good"). 96.55% argued that their equipment either does not have any impairment (72.41%) or that the impairment is not annoying (24.13%). Finally, the majority of workers (58.62%) reported near silence conditions, while 31.03% of them reported normal, non-distracting levels of noise. While these conditions are not comparable to lab setups, we consider them to be sufficiently good to accommodate the requirements of our study.

In the NASA-TLX questionnaire, 34.48% of crowd workers reported low "Mental Demand" and 79.31% low "Physical Demand". "Temporal Demand" was also reported low for the 72.41% of the participants. This low self-reported demand, is reflected also to the majority (55.17%), who reported higher than average "Performance". Nevertheless, the majority of crowd workers (70.11%) reported average to very high amounts of "Effort" while completing the study, which is not reflected on the perceived mental, physical and temporal demand they experienced. It is also not evident on their "Frustration" levels, since the majority (54.02%) reported low levels.

Using Pearson's R, we found the inter-correlations between the different categories of NASA-TLX. We found high correlation between "Physical Demand" and "Mental Demand", but also between "Physical Demand" and "Temporal Demand". Finally, "Frustration" and "Performance" show high correlation between them, which is a reasonable effect.

4.1.5. Identifying Factors Influencing Performance in Mini-PROMS

To better understand factors affecting a participant's performance in Mini-PROMS and therefore their perceptual capabilities on Melody, Tempo, Tuning and Accent, we applied a Multiple Linear Regression, using Ordinary Least Square (OLS) method. We split our analysis based on total score on Mini-PROMS and the individual categories of the test, to study how they are influenced by the rest of the study's categories.

To minimize multi-colinearity between the Independent Variables, we dropped those that showed high correlation between them in our inter-correlation analysis. Analyzing the inter-correlations between all categories, we found similar results to those per part of the study (as analyzed in previous sections). Therefore, NASA-TLX was the only part of the study on Prolific, where high inter-correlation was exhibited between the categories of "Physical Demand" and "Mental Demand", "Physical Demand" and "Temporal Demand", "Frustration" and "Performance". We proceeded to apply OLS, by dropping "Physical Demand" and "Frustration" from the NASA-TLX factors, to decrease colinearity. Correspondingly, for the categorical variables "Occupation" and "Equipment Type", we only used the "Part Time", "Voluntary Work", "Unemployed" and "Headphones", "Laptop Speakers" for each respective variable.

For the total Mini-PROMS score, we found a significant equation [$F_{(19, 67)} = 2.948$, $p < 0.000$, with $R^2 = 0.455$], that shows "Perceptual Abilities" and "Musical Training" from GMSI, affect significantly the dependent variable ($p < 0.05$). For each unit increase reported under the "Perceptual Abilities", a worker showed an increase of 0.2207 point in the total score, while in "Musical Training", it resulted to a 0.2417 increase.

Running the regression for the "Melody" of Mini-PROMS [$F_{(19, 67)} = 1.898$, $p = 0.0289$, with $R^2 = 0.350$], we found that their "Occupation" status affected the dependent variable significantly ($p < 0.05$). Their "Part Time" employment seemed to negatively influence their performance in "Melody" test, by -1.2234 points. On the other hand, "Perceptual Abilities" and "Musical Training" from GMSI also affected significantly their performance ($p < 0.05$), increasing it by 0.0827 and 0.0570 points, respectively. Their "Singing Abilities" though, seemed to significantly influence their performance but negatively, where every reported increase on those abilities, resulted to a decrease of -0.0672 point.

The significant regression equation that was found for the "Tuning" category [$F_{(19, 67)} = 2.301$, $p = 0.006$, with $R^2 = 0.395$] showed that their "Occupation" status was yet again affecting their performance significantly ($p < 0.05$). Those who reported "Unemployed" showed an increase in their performance by 0.9205. Finally, "Musical Training" appears to be another significant factor to their performance in this particular audio test. Each unit increase in the category, resulted in a 0.0709 increase in their performance.

For "Accent", the regression [$F_{(19, 67)} = 2.580$, $p = 0.002$, with $R^2 = 0.422$], showed that the "Temporal Demand" the participants experienced, alongside their "Occupation" and "Musical Training", influenced significantly their performance in this test. An increase in "Temporal Demand" resulted in decrease by -0.0851 point and in "Musical Training", an increase by a 0.0701 point. "Part Time" occupation is negatively associated with their performance here, leading to a decrease of -1.2801 points.

Finally, for the "Tempo" category of Mini-PROMS, we couldn't find a significant model by applying OLS.

4.2. Amazon Mechanical Turk

Of the 100 workers recruited from Amazon Mechanical Turk (MTurk), 9 of them failed at least one attention check question(s); 7 of them provided invalid/none inputs. After excluding these

TABLE 5 | MTurk participant demographics.

	Variables	Statistics
Age (years)	Range	18–65+
	Majority	26–35 (52.38%)
Occupation	Full-time	71
	Part-time	9
	Unemployed	3
	Retired	1
Education	Associate degree	6
	Bachelor's degree	44
	Doctorate degree	2
	High school/HED	11
	Master's degree	10
	Professional degree	0
	Some college, no diploma	8
	Some high school, no diploma	1
	Technical/trade/vocational training	2

16 invalid submissions, we have 84 valid submissions from 84 unique workers.

4.2.1. Worker Demographics

We also conducted the same study on Amazon MTurk, in order to see if we can observe similar trends as shown in the last section also on a platform different than Prolific. We gathered 84 crowd workers who provided valid submissions. As seen in **Table 5**, the age range was between 18 and above 65, while the majority was between 26-35 (52.38%), a relatively older pool compared to the Prolific's one. The majority of them were employed (95.23%), with the 88.75% of them full-time. Most of the participants hold a degree (86.90%), with Bachelor's being the most common (60.27%). Finally, the vast majority of the participants, report the United States of America (89.28%) as their residence, with the rest being spread between Brazil (3), India (3), United Kingdom (1), Netherlands (1), and Italy (1).

Apart from education, we see a clear difference between the participants from the two platforms on the age, occupation and country of residence categories. In this study, most of the crowd workers from MTurk are older than those on Prolific, employed and residing in USA.

4.2.2. Results on Worker Music Sophistication

In **Table 6**, we summarize the results of the GMSI questionnaire, regarding the workers on MTurk. As described in Section 4.1.2, we compare the results on this platform, with the results of the original GMSI study (Müllensiefen et al., 2014).

Comparing our collected data to the original GMSI study, we find that the crowd workers of MTurk exhibit a strongly lower overall music sophistication, at the bottom 32% of the original study. They also score low in all sub-categories, with Musical Training being the only category comparing higher to the 37% of the original study's population.

TABLE 6 | GMSI range, median, mean, and standard deviation.

	Range	Median	Mean	Standard deviation (1σ)
Active engagement	12–46	32	30.57	7.92
Perceptual abilities	18–47	32.5	32.82	5.92
Musical training	7–43	23	21.80	9.40
Singing abilities	9–45	32.5	28.29	8.12
Emotions	7–41	30.5	30.34	5.35
General music sophistication	29–113	75	72.19	18.15

TABLE 7 | Intercorrelations (Spearman's rank) of subscales of GMSI scores.

	Active engagement	Perceptual abilities	Musical training	Emotions	Singing abilities
Active engagement	1.000				
Perceptual abilities	0.232*	1.000			
Musical training	**0.595***	0.263*	1.000		
Emotions	0.213	0.471*	-0.052	1.000	
Singing abilities	**0.637***	0.340*	**0.552***	0.223*	1.000

Statistical significance ($p < 0.05$) is marked using an asterisk (*). Values in bold indicate intercorrelations higher than 0.5.

An extremely high number of participants (40.47%), reported having "Absolute Pitch", which is a highly unlikely portion of the sample, as discussed before. Only 9 of them reported adequate formal musical training, which can indicate a general misconception on the entailing traits of such a phenomenon. The reports are much higher than those on Prolific.

With a quick glance at the values on **Table 7**, we see that they indicate skewness on the distributions of each category. When running the Shapiro-Wilk normality test (Hanusz et al., 2016), we found that all distributions, except that of "Perceptual Abilities", are non-normal. For that reason, we used Spearman's ranked test to calculate the correlation coefficients between the GMSI sub-categories.

We find that "Active Engagement", "Musical Training" and "Singing Abilities" are highly correlated with each other. The positive high correlation between these categories, indicates that crowd workers on MTurk report similarly their aptitude on those GMSI categories. Notably, although not particularly high, there is certainly a positive correlation between self-reported "Perceptual Abilities" and the extent of "Emotions" these crowd workers experience when listening to music ($R = 0.471$).

4.2.3. Results on Objective Music Perception Skills

Table 8 shows the results of MTurk's crowd workers on Mini-PROMS test. The mean overall score shows that the average participant in our sample pool, had lower than "Basic" music perception skills overall ($\mu = 15.2$ 42.2% avg. accuracy). This performance is much lower than both the original Mini-PROMS work (Zentner and Strauss, 2017) and the results we retrieved from Prolific. In the Top 10% of the highest performant crowd workers, we see that they score from 61.1% up to 81.94%, scoring from "Good" to "Outstanding", based on the Mini-PROMS scale.

TABLE 8 | Mini-PROMS range, median, mean, and standard deviation.

	Range	Median	Mean	Standard deviation (1σ)
Melody	1–8	4.25	4.22	1.56
Tuning	1–7.5	3	3.2	1.33
Accent	0–7.5	4	4.04	1.42
Tempo	1–8	3.5	3.75	1.64
Mini-PROMS	6.5–29.5	14.5	15.2	4.77

Per individual categories, we see that the highest performance that the crowd workers achieved, didn't reach the max of the Mini-PROMS scale of every category except "Tempo". The avg. accuracy on the "Melody" category, reached 42.2% (original study: 64.3%, Prolific: 49.77%), while on the "Tuning" category, the avg. accuracy was 40% (original study: 68%, Prolific: 52.73%). The participants from MTurk, were able to detect changes on "Accent" features with avg. accuracy of 40.4% (original study: 61.5%, Prolific: 51.95%), while they scored avg. accuracy 46.87% on "Tempo" (original study: 81.25%, Prolific: 64.3%).

Running the Shapiro-Wilk normality test on each Mini-PROMS' category, we find that only the "Melody" one is Gaussian. We used once again Spearman's rank method to calculate the correlation coefficients per category. We found that "Tempo" is highly correlated with "Melody", while "Tuning" is with "Accent". These results are not in line with the original PROMS study (Law and Zentner, 2012), but we observe relatively strong correlation between "Tuning"-"Melody" and "Tuning"-"Tempo", which fall into the PROMS categories of "Sound perception" and "Sensory" skills, respectively.

4.2.4. Post-task Survey: Equipment and Cognitive Workload

The majority of the participants from MTurk used headphones to perform Mini-PROMS (64.28%), while 20.23% used earphones and 15.46% used the speakers of their laptops. Most participants described the condition of their equipment as "Excellent" or "Good", while one reported it as "Fair". The majority (66.66%) of the crowd workers, reported any impairment of their equipment as "Impairceptible", with 15.47% of them describing it as "Perceptible but not annoying". The rest of the workers reported various degrees of annoying impairments. Finally, 65.47% of the crowd workers performed Mini-PROMS with near silence environmental conditions, while 19.04% reported extreme levels of noise around them. None of the distributions passed the Shapiro-Wilk normality test for each equipment-related category. Running Spearman's rank method, we found no notable correlation between the categories.

In the NASA-TLX questionnaire, 29.76% of crowd workers reported average "Mental Demand", with 10.71 and 15.67% reporting low or very high mental strain, respectively. 46.43% reported low "Physical Demand" with 36.9% of the total not feeling rushed while performing the study. 27.38% reported average "Performance", with 22.61% describing their performance as successful. The majority of participants were divided between reporting high effort (27.38%) or moderate difficulty (27.38%). Finally, 34.52% of the crowd workers felt little to no frustration with 20.24% reporting moderate levels.

Using Spearman's rank, we found that "Frustration" is highly correlated with "Physical Demand", "Temporal Demand" and "Performance". This shows that the more physical strain and hurried they felt, combined with feelings of failing the task at hand, increased their frustration with the study.

4.2.5. Identifying Factors Influencing Performance in Mini-PROMS

Following the analysis on the results from Prolific, we applied Multiple Linear Regression on the Mini-PROMS categories, using the Ordinary Least Square (OLS) method. For the total Mini-PROMS score as the dependent variable [$F_{(18,\,65)} = 4.742$, $p < 0.000$, with $R^2 = 0.567$], we found that only the "Perceptual Skills" from GMSI and the "Physical Demand" category from NASA-TLX, affect significantly the dependent variable ($p < 0.05$). For each extra point reported under the "Perceptual Skills", a worker showed an increase of 0.2 points in the total score. On the other hand, a single extra point toward "Very demanding" on the "Physical Demand" category, resulted on a -0.3 decrease of total performance by the worker.

Running the regression for the "Melody" of Mini-PROMS [$F_{(18,\,65)} = 3.443$, $p < 0.000$, with $R^2 = 0.488$], we found that "Physical Demand" category from NASA-TLX affected the dependent variable the most ($p < 0.05$). The effect is negative toward the performance on "Melody", where each point increase on "Physical Demand" translated to -0.12 point decrease of performance.

The significant regression equation that was found for the "Tuning" category [$F_{(18,\,65)} = 1.849$, $p = 0.0376$, with $R^2 = 0.339$] showed that the most significant factor was yet again the "Physical Demand". The more physically demanding the study was perceived, it influenced the final score on "Tuning" by -0.09.

For "Accent", the regression [$F_{(18,\,65)} = 2.130$, $p = 0.0141$, with $R^2 = 0.371$], showed that the "Type" of audio equipment and its "Impairment" affected the workers' performance the most. The "Laptop Speakers" seemed to influenced positively their performance by 1.2193 points, while the less perceptible an "Impairment" was, it was increasing their performance by 0.448 point.

Finally, for the "Tempo" category of Mini-PROMS, we found a significant regression equation [$F_{(18,\,65)} = 4.502$, $p < 0.000$, with $R^2 = 0.555$] that shows that "Physical Demand" and "Perceptual Abilities" influenced the performance on the category most significantly. While an increase in "Physical Demand" decreased the performance by -0.10 point, an increase in the self-reported "Perceptual Abilities" showed an increase of performance on "Tempo" by 0.08.

4.3. In Search of Musical Sleepers

Having analyzed each component of our study and using OLS to understand how individual factors could have influenced the workers' performance on the perception skills of Mini-PROMS, we were still interested to investigate how the highly perceptive workers are distributed based on quantifiable expertise. Musical training is an element that can be quantified by questions on credentials, years of education etc, all components that can be retrieved by the respective category in GMSI. It is an attribute that we experts show high proficiency and that a platform could potentially easily store and iterate per worker's profile.

In this study, following the original studies of PROMS (Law and Zentner, 2012) and Mini-PROMS (Zentner and Strauss, 2017), we make the comparisons of levels of Musical Training, against the performance on the categories of Mini-PROMS. Taking a step further, we used as baselines the amount of "Musical Training" that 50% of the original GMSI's population exhibited (27) and the lowest bound of "Excellent" performance (63.98%) on perception skills, as established for Mini-PROMS. We make use of the terms "Musical Sleepers", to label those who exhibit high performance but reported low training and "Sleeping Musicians", those who reported extensive training but performed poorly, both terms from Law and Zentner (2012) and Zentner and Strauss (2017).

Figure 3 shows a scatter plot per platform, that shows how participants are distributed based on their performance and "Musical Training". We witness that on both platforms, there is a high number of crowd workers who reported low "Musical Training" [below 50% of original GMSI study in Müllensiefen et al. (2014)] and had relatively low performance in the Mini-PROMS tests. This is to be expected, due to the nature of the domain and the niche skills that are required.

The attention is naturally drawn to the "Musical Sleepers"; a portion of the population that can exhibit relatively high music perception skills, but did not have adequate education. Few people would follow any form of dedicated music studies, making it even more difficult to find them on a crowdsourcing platform. With low expertise being the norm, finding crowd workers with high, untrained, auditory skills among them, is a rare phenomenon that could greatly benefit systems who would make use of such skills. In the case of Prolific, we witness "Musical Sleepers" in a higher number compared to Amazon Mechanical Turk. We cannot draw platform-based conclusions though, since our participant pool was quite small relatively to the actual population of each platform. The presence of these workers is very encouraging, as it shows that it is possible to deploy advanced music analysis tasks on microtask platforms and finding high-value contributors.

In our study, participants from Amazon Mechanical Turk, generally reached lower performance compared to the ones from Prolific. This is an outcome also evident on the high number of "Sleeping Musicians" on MTurk, compared to the smaller portion of the total Prolific participants. These workers reported relatively high musical training, but performed lower than expected from a person of their expertise.

5. DISCUSSION

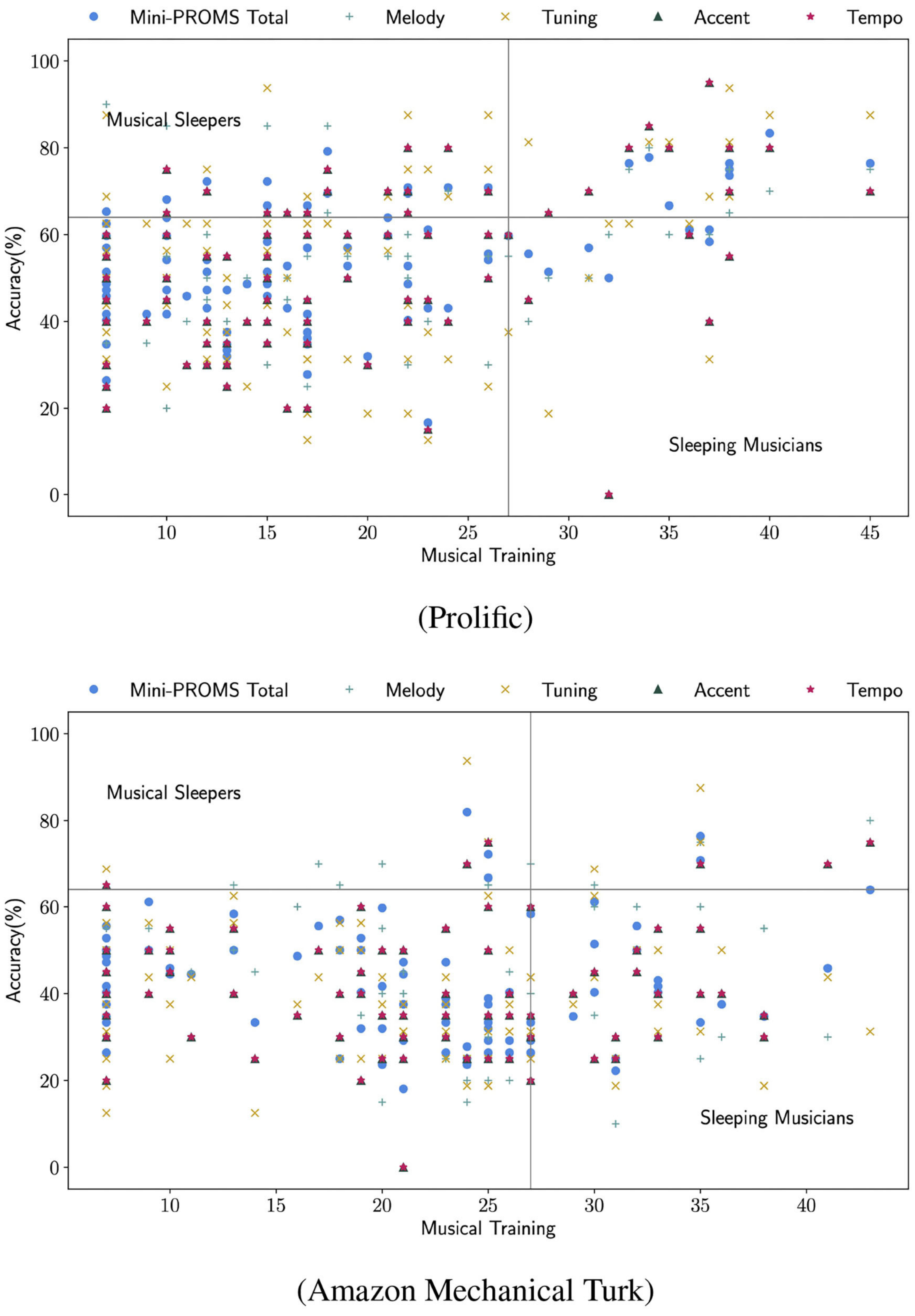
In this study, we extensively measure the musical sophistication and music perception skills of crowd workers on Prolific and Amazon Mechanical Turk. We show that on both platforms, the self-reported music sophistication of crowd workers is below that of the general population and that formally-trained workers are rare. Nevertheless, we found surprisingly refined and diverse music perception skills amongst the top performers per platform. These skills though cannot accurately and easily be predicted by questions.

5.1. On Music Perceptual Skills and Predictors

Workers on both platforms exhibited quite diverse set of music perception skills. Among the high performant ones, we found evidence that supports the existence of workers with high accuracy and little to no formal training, namely "Musical Sleepers", indicating the prospect of high-quality annotations by non-experts on these platforms. Predicting these skills though, can prove far from trivial. To promote reproducibility of our results, we made use of established tools to retrieve domain sophistication (Müllensiefen et al., 2014), perceptual skills (Zentner and Strauss, 2017), perceived workload (Hart and Staveland, 1988), equipment condition (ITU-R BS, 2003) and ambient noise levels (Beach et al., 2012).

In an analysis of workers' reports on other parts of the study, we found per platform, different factors that significantly correlated to their performance. "Musical Training", a type of expertise that could be thought as a strong indicator of a worker's perceptual skills, showed low significance on the performance of Amazon Mechanical Turk workers. These findings, alongside the high number of "Sleeping Musicians" among the participants from Amazon Mechanical Turk, indicate a notable difference between their reported knowledge and their quantified perceptual skills. On the other hand though, the self-reported "Perceptual Abilities" proved a reliable factor of MTurk workers, as they were significantly related to their performance on Mini-PROMS. This is in contrast to the reported "Perceptual Abilities" of Prolific's workers, which did not significantly correlate to their performance. Aspects of perceived task workload though, as retrieved from NASA-TLX, seemed to significantly correlate on categories of the Mini-PROMS test, on both platforms. Finally, while demographic data appear relevant to aspects of the performance of workers on Prolific, on MTurk equipment showed to play a more important role on the "Accent" test of Mini-PROMS.

The "Active Engagement" category of GMSI, which indicates to what extent a person engages with music as a hobby (frequenting online forums, buying music albums, etc.), did not show any significant correlation to the measured music perception skills of the participants on both platforms. That shows that we cannot reliably use such questions, to infer the skills of the worker; the time/effort spent listening to or discussing about music, can be indifferent of the range of their skills. The same applies to the "Emotions" category, where participants report their emotional response to music. This indicates that music could still evoke emotions to people, even without them perceiving its structural elements.



(Prolific)

(Amazon Mechanical Turk)

FIGURE 3 | Musical Training (GMSI) and Performance on Mini-PROMS (acc%).

5.2. Implications for Design

Self-reported Musical Sophistication. The musical sophistication assessments (GMSI) is a useful tool to evaluate workers' capability in completing music-related tasks. It is however a lengthy questionnaire, which could result in extra cost and worse worker engagement. Reducing the number of question is possible, but with implication in terms of test reliability. For instance, the subscale of Musical Training is positively correlated to their actual music perception skills (and the correlation coefficient is higher than the general GMSI). As music perception skills are of primary relevance when executing music-related tasks, we suggest that in future task design, requesters could consider using the subscale of musical training which only contains 7 items. This could be complemented with novel methods to effectively and precisely predict worker performance to further facilitate task scheduling and assignment.

Music Perception Skill Assessment. The Mini-PROMS tool appears to be an effective mean to evaluate worker quality in terms of music skills. Yet, it suffers from the same overhead issues of GMSI. In this case, we suggest to use PROMS or Mini-PROMS as a qualification test, possibly featured by crowdsourcing platforms. Workers could use this test to get the corresponding qualification, to obtain the opportunities to access more tasks, and earn more rewards.

Music Annotation and Analysis Tasks. The results of this study indicate that knowledge- and skill-intensive musical tasks could be deployed on microtasks crowdsourcing platforms, with good expectations in terms of availability of skilled workers. However, performance on different skills (Melody, Tuning, Accent, and Tempo) appears to be unevenly distributed. We therefore recommend to analyse the capabilities of the selected crowd and tailor the design of advanced music annotation and analysis tasks to precise music perception skills.

5.3. Limitations and Future Work

A main limitation of our study is concerned with the size of the tested population. While we employed workers from two different platforms, our results cannot be generalized per platform. A larger participation pool could potentially aid the generalisability of our findings and lead to more fine-grained insights. Even though our results are based on a population of crowd workers that have received less formal musical training than the average population used in similar studies (Müllensiefen et al., 2014) the use of standardized and validated tests, lend confidence to the reliability of our findings.

Another potential confounding factor in our study is the motivation for participation. We attracted crowd workers using monetary rewards, while in other studies people voluntarily performed their test (e.g., BBC's main Science webpage, Müllensiefen et al., 2014). Such a difference could also explain the differences in observed distributions (musical training and perception skills). However, monetary incentives are a feature of crowdsourcing markets, which makes them appealing in terms of work capacity and likelihood of speedy completion. In that respect, our findings are very encouraging, as they show the availability of both musically educated and/or naturally skilled workers that could take on musically complex tasks.

As demonstrated in our results, workers who perform well in a certain perception category (e.g., "Melody") do not perform equally well in another (e.g., "Tempo"). In future studies, we encourage the use of perception tests, adjusted and adapted for the specific music task at hand by using the appropriate categories, to accurately select potentially highly performing workers.

In our analysis, we currently made use of Ordinary Least Square Regression to identify factors are associated with the workers' performance on Mini-PROMS. Although this method gave us some first insights, further studies are needed to expand our pool of crowd workers and use other models that can help us find predictors of perceptual skills of workers accurately. This could assist in designing appropriate task assignment methods, to increase the efficiency and effectiveness of crowdsourcing systems that make use of such skills.

In this study, we utilized standardized tools to capture domain-specific characteristics of the workers of a specific platform. Comparing results from their self-reported "connection" to the domain, with those from actively testing their skills, can paint a clear picture of the workers' demographics on a specific domain. While this work is specific to the music domain, we believe that similar workflows can be utilized to study the characteristics of workers on other domains. This holds especially true, as crowdsourcing platforms have diverse user-bases and direct comparisons cannot safely be drawn to studies with highly controlled population samples.

6. CONCLUSION

In this paper, we have presented a study exploring the prevalence and distribution of music perception skills of the general crowd in the open crowdsourcing marketplace of Prolific and Amazon Mechanical Turk. We measured and compared self-reported musical sophistication and active music perception skills of crowd workers by leveraging the established GMSI questionnaire and Mini-PROMS audio-based test, respectively. Our analysis shows that self-reported musical sophistication of crowd workers is generally below that of the general population and the majority of them have not received any form of formal training. We observed differences in the two participant pools, on both their performance and factors which are significantly correlated to it. Nevertheless, we identified the presence of *musical sleepers* on both platforms. Moreover, our analysis shows worker accessibility to adequate equipment. Together, these findings indicate the possibility of further increasing the adoption of crowdsourcing as a viable means to perform complex music-related tasks.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Delft University of Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

IS, SQ, AB, and CL contributed to conception and design of the study. IS implemented the interfaces and online system and wrote the first draft of the manuscript. IS and SQ performed the statistical analysis. All authors wrote sections of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

Baharloo, S., Johnston, P. A., Service, S. K., Gitschier, J., and Freimer, N. B. (1998). Absolute pitch: an approach for identification of genetic and nongenetic components. *Am. J. Hum. Genet*. 62, 224–231. doi: 10.1086/301704

Baharloo, S., Service, S. K., Risch, N., Gitschier, J., and Freimer, N. B. (2000). Familial aggregation of absolute pitch. *Am. J. Hum. Genet*. 67, 755–758. doi: 10.1086/303057

Beach, E. F., Williams, W., and Gilliver, M. (2012). The objective-subjective assessment of noise: young adults can estimate loudness of events and lifestyle noise. *Int. J. Audiol*. 51, 444–449. doi: 10.3109/14992027.2012.658971

Burkhard, A., Elmer, S., and Jäncke, L. (2019). Early tone categorization in absolute pitch musicians is subserved by the right-sided perisylvian brain. *Sci. Rep*. 9, 1–14. doi: 10.1038/s41598-018-38273-0

De Prisco, R., Esposito, A., Lettieri, N., Malandrino, D., Pirozzi, D., Zaccagnino, G., et al. (2017). "Music plagiarism at a glance: metrics of similarity and visualizations," in *2017 21st International Conference Information Visualisation (IV)* (London: IEEE), 410–415. doi: 10.1109/iV.2017.49

De Prisco, R., Lettieri, N., Malandrino, D., Pirozzi, D., Zaccagnino, G., and Zaccagnino, R. (2016). "Visualization of music plagiarism: analysis and evaluation," in *2016 20th International Conference Information Visualisation (IV)* (Lisbon: IEEE), 177–182. doi: 10.1109/IV.2016.56

Degrave, P., and Dedonder, J. (2019). A French translation of the goldsmiths musical sophistication index, an instrument to assess self-reported musical skills, abilities and behaviours. *J. N. Music Res*. 48, 138–144. doi: 10.1080/09298215.2018.1499779

Demorest, S. M., Morrison, S. J., Jungbluth, D., and Beken, M. N. (2008). Lost in translation: an enculturation effect in music memory performance. *Music Percept*. 25, 213–223. doi: 10.1525/mp.2008.25.3.213

Gadiraju, U., Möller, S., Nöllenburg, M., Saupe, D., Egger-Lampl, S., Archambault, D., et al. (2017). "Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd," in *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments* (Dagstuhl Castle: Springer), 6–26. doi: 10.1007/978-3-319-66435-4_2

Gingras, B., Honing, H., Peretz, I., Trainor, L. J., and Fisher, S. E. (2015). Defining the biological bases of individual differences in musicality. *Philos. Trans. R. Soc. B Biol. Sci*. 370, 20140092. doi: 10.1098/rstb.2014.0092

Hannon, E. E., and Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends Cogn. Sci*. 11, 466–472. doi: 10.1016/j.tics.2007.08.008

Hanusz, Z., Tarasinska, J., and Zielinski, W. (2016). Shapiro-wilk test with known mean. *REVSTAT Stat. J*. 14, 89–100. doi: 10.57805/revstat.v14i1.180

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (task load index): results of empirical and theoretical research," in *Advances in Psychology, Vol. 52*, eds P. A. Hancock and N. Meshkati (North-Holland: Elsevier), 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hyde, K. L., and Peretz, I. (2004). Brains that are out of tune but in time. *Psychol. Sci*. 15, 356–360. doi: 10.1111/j.0956-7976.2004.00683.x

ITU-R BS (2003). *General Methods for the Subjective Assessment of Sound Quality*.

Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert scale: explored and explained. *Brit. J. Appl. Sci. Technol*. 7, 396. doi: 10.9734/BJAST/2015/14975

Koelsch, S., Schulze, K., Sammler, D., Fritz, T., Müller, K., and Gruber, O. (2009). Functional architecture of verbal and tonal working memory: an fMRI study. *Hum. Brain Mapp*. 30, 859–873. doi: 10.1002/hbm.20550

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol*. 77, 1121. doi: 10.1037/0022-3514.77.6.1121

Law, E. L., Von Ahn, L., Dannenberg, R. B., and Crawford, M. (2007). "Tagatune: a game for music and sound annotation," in *ISMIR* (Vienna), 2.

Law, L. N., and Zentner, M. (2012). Assessing musical abilities objectively: construction and validation of the profile of music perception skills. *PLoS ONE* 7, e52508. doi: 10.1371/journal.pone.0052508

Lee, J. H. (2010). "Crowdsourcing music similarity judgments using mechanical turk," in *ISMIR* (Utrecht), 183–188.

Lee, J. H., Hill, T., and Work, L. (2012). "What does music mood mean for real users?" in *Proceedings of the 2012 IConference, iConference '12* (New York, NY: Association for Computing Machinery), 112–119. doi: 10.1145/2132176.2132191

Lee, J. H., and Hu, X. (2012). "Generating ground truth for music mood classification using mechanical turk," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries* (Washington, DC), 129–138. doi: 10.1145/2232817.2232842

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lima, C. F., Correia, A. I., Müllensiefen, D., and Castro, S. L. (2020). Goldsmiths musical sophistication index (gold-MSI): Portuguese version and associations with socio-demographic factors, personality and music preferences. *Psychol. Music* 48, 376–388. doi: 10.1177/0305735618801997

Lin, H.-R., Kopiez, R., Müllensiefen, D., and Wolf, A. (2021). The Chinese version of the gold-MSI: adaptation and validation of an inventory for the measurement of musical sophistication in a Taiwanese sample. *Music. Sci*. 25, 226–251. doi: 10.1177/1029864919871987

Malandrino, D., Pirozzi, D., Zaccagnino, G., and Zaccagnino, R. (2015). "A color-based visualization approach to understand harmonic structures of musical compositions," in *2015 19th International Conference on Information Visualisation* (Barcelona: IEEE), 56–61. doi: 10.1109/iV.2015.21

Mandel, M. I., Eck, D., and Bengio, Y. (2010). *Learning Tags That Vary Within a Song*. ISMIR, Utrecht.

Mankel, K., and Bidelman, G. M. (2018). Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proc. Natl. Acad. Sci. U.S.A*. 115, 13129–13134. doi: 10.1073/pnas.1811793115

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE* 9, e89642. doi: 10.1371/journal.pone.0089642

Oh, J., and Wang, G. (2012). "Evaluating crowdsourcing through amazon mechanical turk as a technique for conducting music perception experiments," in *Proceedings of the 12th International Conference on Music Perception and Cognition* (Thessaloniki), 1–6.

Oosterman, J., Yang, J., Bozzon, A., Aroyo, L., and Houben, G.-J. (2015). On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks. *Comput. Netw.* 90, 133–149. doi: 10.1016/j.comnet.2015.07.008

Peretz, I., and Hyde, K. L. (2003). What is specific to music processing? Insights from congenital amusia. *Trends Cogn. Sci.* 7, 362–367. doi: 10.1016/S1364-6613(03)00150-5

Reymore, L., and Hansen, N. C. (2020). A theory of instrument-specific absolute pitch. *Front. Psychol.* 11, 2801. doi: 10.3389/fpsyg.2020.560877

Samiotis, I. P., Qiu, S., Lofi, C., Yang, J., Gadiraju, U., and Bozzon, A. (2021). "Exploring the music perception skills of crowd workers," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 108–119.

Samiotis, I. P., Qiu, S., Mauri, A., Liem, C. C. S., Lofi, C., and Bozzon, A. (2020). "Microtask crowdsourcing for music score transcriptions: an experiment with error detection," in *Proceedings of the 21st International Society for Music Information Retrieval Conference* (Montréal, QC).

Schaal, N. K., Bauer, A.-K. R., and Müllensiefen, D. (2014). Der gold-MSI: replikation und validierung eines fragebogeninstrumentes zur messung musikalischer erfahrenheit anhand einer deutschen stichprobe. *Music. Sci.* 18, 423–447. doi: 10.1177/1029864914541851

Sorokin, A., and Forsyth, D. (2008). "Utility data annotation with amazon mechanical turk," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (Anchorage, AK: IEEE), 1–8. doi: 10.1109/CVPRW.2008.4562953

Speck, J. A., Schmidt, E. M., Morton, B. G., and Kim, Y. E. (2011). "A comparative study of collaborative vs. traditional musical mood annotation," in *ISMIR* (Miami, FL: Citeseer), 549–554.

Totterdell, P., and Niven, K. (2014). *Workplace Moods and Emotions: A Review of Research*. Charleston, SC: Createspace Independent Publishing.

Ullén, F., Mosing, M. A., Holm, L., Eriksson, H., and Madison, G. (2014). Psychometric properties and heritability of a new online test for musicality, the swedish musical discrimination test. *Pers. Individ. Diff.* 63, 87–93. doi: 10.1016/j.paid.2014.01.057

Urbano, J., Morato, J., Marrero, M., and Martin, D. (2010). "Crowdsourcing preference judgments for evaluation of music similarity tasks," in *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation* (New York, NY: ACM), 9–16.

Werner, P. D., Swope, A. J., and Heide, F. J. (2006). The music experience questionnaire: development and correlates. *J. Psychol.* 140, 329–345. doi: 10.3200/JRLP.140.4.329-345

Zentner, M., and Strauss, H. (2017). Assessing musical ability quickly and objectively: development and validation of the short-proms and the mini-proms. *Ann. N. Y. Acad. Sci.* 1400, 33–45. doi: 10.1111/nyas.13410

Zhuang, M., and Gadiraju, U. (2019). "In what mood are you today? An analysis of crowd workers' mood, performance and engagement," in *Proceedings of the 10th ACM Conference on Web Science*, 373–382. doi: 10.1145/3292522.3326010

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 29 June 2022
doi: 10.3389/frai.2022.848056

Improving Crowdsourcing-Based Image Classification Through Expanded Input Elicitation and Machine Learning

Romena Yasmin[1]*, *Md Mahmudulla Hassan*[2], *Joshua T. Grassel*[1], *Harika Bhogaraju*[1], *Adolfo R. Escobedo*[1] *and Olac Fuentes*[2]

[1] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, United States, [2] Department of Computer Science, University of Texas at El Paso, El Paso, TX, United States

OPEN ACCESS

Edited by:
Matt Lease,
University of Texas at Austin,
United States

Reviewed by:
Valentina Poggioni,
University of Perugia, Italy
Evgenia Christoforou,
CYENS—Centre of Excellence,
Cyprus

***Correspondence:**
Romena Yasmin
ryasmin@asu.edu

Specialty section:
This article was submitted to
Machine Learning and Artificial
Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

Received: 03 January 2022
Accepted: 24 May 2022
Published: 29 June 2022

Citation:
Yasmin R, Hassan MM, Grassel JT,
Bhogaraju H, Escobedo AR and
Fuentes O (2022) Improving
Crowdsourcing-Based Image
Classification Through Expanded Input
Elicitation and Machine Learning.
Front. Artif. Intell. 5:848056.
doi: 10.3389/frai.2022.848056

This work investigates how different forms of input elicitation obtained from crowdsourcing can be utilized to improve the quality of inferred labels for image classification tasks, where an image must be labeled as either positive or negative depending on the presence/absence of a specified object. Five types of input elicitation methods are tested: binary classification (positive or negative); the (x, y)-coordinate of the position participants believe a target object is located; level of confidence in binary response (on a scale from 0 to 100%); what participants believe the majority of the other participants' binary classification is; and participant's perceived difficulty level of the task (on a discrete scale). We design two crowdsourcing studies to test the performance of a variety of input elicitation methods and utilize data from over 300 participants. Various existing voting and machine learning (ML) methods are applied to make the best use of these inputs. In an effort to assess their performance on classification tasks of varying difficulty, a systematic synthetic image generation process is developed. Each generated image combines items from the *MPEG-7 Core Experiment CE-Shape-1 Test Set* into a single image using multiple parameters (e.g., density, transparency, etc.) and may or may not contain a target object. The difficulty of these images is validated by the performance of an automated image classification method. Experiment results suggest that more accurate results can be achieved with smaller training datasets when both the crowdsourced binary classification labels and the average of the self-reported confidence values in these labels are used as features for the ML classifiers. Moreover, when a relatively larger properly annotated dataset is available, in some cases augmenting these ML algorithms with the results (i.e., probability of outcome) from an automated classifier can achieve even higher performance than what can be obtained by using any one of the individual classifiers. Lastly, supplementary analysis of the collected data demonstrates that other performance metrics of interest, namely reduced false-negative rates, can be prioritized through special modifications of the proposed aggregation methods.

Keywords: machine learning, input elicitations, crowdsourcing, human computation, image classification

1. INTRODUCTION

In recent years, computer vision approaches based on machine learning (ML) and, in particular, those based on deep convolutional neural networks have demonstrated significant performance improvements over conventional approaches for image classification and annotation (Krizhevsky et al., 2012; Tan and Le, 2019; Zhai et al., 2021). However, these algorithms generally require a large and diverse set of annotated data to generate accurate classifications. Large amounts of annotated data are not always available, especially for tasks where producing high-quality meta-data is costly, such as image-based medical diagnosis (Cheplygina et al., 2019), pattern recognition in geospatial remote sensing data (Rasp et al., 2020; Stevens et al., 2020), etc. In addition, ML algorithms are often sensitive to perturbations in the data for complex visual tasks, that to some extent are even difficult for humans, such as object detection in cluttered backgrounds and detection of adversarial examples (McDaniel et al., 2016; Papernot et al., 2016), due to the high dimensionality and variability of the feature space of the images.

Crowdsourcing has received significant attention in various domain-specific applications as a complementary approach for image classification. Its growth has been accompanied and propelled by the emergence of online crowdsourcing platforms (e.g., Amazon Mechanical Turk, Prolific), which are widely employed to recruit and compensate human participants for annotating and classifying data that are difficult for machine-only approaches. In general, crowdsourcing works by leveraging the concept of the "wisdom of the crowd" (Surowiecki, 2005), with which the judgments or predictions of multiple participants are aggregated to sift out noise and to better approximate a ground truth (Yi et al., 2012). Numerous studies over the last decade have established that, under the right circumstances and with the proper aggregation methods, the collective judgment of multiple non-experts is uncontroversially more accurate than those from almost any individual, including well-informed experts. This concept of using groups to make collective decisions has been successfully applied to a number of visual tasks ranging from simple classification and annotation (Russakovsky et al., 2015) to complex real-world applications, including assessment of damages caused by natural disasters (Barrington et al., 2012) and segmentation of biomedical images for diagnostic purposes (Gurari et al., 2015).

Although ML methods have been shown to perform exceedingly well in various classification tasks, these outcomes typically depend on relatively large datasets (Hsing et al., 2018). However, high amounts of richly annotated data are inaccessible in various situations and/or obtaining them is prohibitively costly. Yet in such situations where less data is available, ML methods provide a natural mechanism for incorporating multiple forms of crowdsourced inputs, since they are tailor-made for classification based on input features. Previous works have tended to use a single form of input (i.e., mostly binary classification labels provided by participants) as a feature for ML algorithms on visual classification tasks. However, the vast majority have overlooked other inputs that can be elicitated from the crowd. Formal studies on the merits and potential impacts of different types of elicited inputs are also lacking. This work investigates how the performance of crowdsourcing-based voting and ML methods for image classification tasks can be improved using a variety of inputs. In summary, the contributions of this work stem from the following objectives:

- Analyze the reliability and accuracy of different ML classifiers on visual screening tasks when different forms of elicited inputs are used as features.
- Evaluate the performance of the classifiers with these additional features on both balanced and imbalanced datasets—i.e., sets of images with equal and unequal proportions, respectively, of positive to negative images—of varying difficulty.
- Introduce supplementary crowdsourcing-based methods to prioritize other performance metrics of interest, namely reduced false-negative and false-positive rates.
- Analyze the performance of the crowdsourcing-based ML classifiers when outputs of an automated classifier trained on large annotated datasets are used as an additional feature.

To pursue these objectives, we design a number of experiments that elicit a diversity of inputs on each classification task: binary classification (1 = positive or 0 = negative); the (x, y)-coordinate of the target object's location; level of confidence in the binary response (on a scale from 0 to 100%); guess of what the majority of participants' binary classification is on the same task; and level of the perceived difficulty of the binary classification task (on a discrete scale). To harness the benefits of both collective human intelligence and machine intelligence, we use the elicited inputs as features for ML algorithms. The results indicate that integrating diverse forms of input elicitation, including self-reported confidence values, can improve the accuracy and efficiency of crowdsourced computation. As an additional contribution, we develop an automated image classification method based on the ResNet-50 neural network architecture (He et al., 2015) by training it on multiple datasets of sizes ranging from 10 k to 90 k image samples. The outputs of this automated classifier are used as additional features within the crowdsourcing-based ML algorithms. These additional results demonstrate that this hybrid image classification approach can provide more accurate predictions, especially for relatively larger datasets, than what is possible by either of the two stand-alone approaches.

Before proceeding, it is pertinent to mention that an earlier, shorter version of this work and a subset of its results appeared in Yasmin et al. (2021) and were presented at the 9th AAAI Conference on Human Computation and Crowdsourcing. That earlier conference paper considered only a subset of the crowdsourcing-based ML algorithms featured herein and that smaller selection was implemented only on balanced datasets. This present work also introduces a hybrid image classification approach, and it incorporates additional descriptions, crowdsourcing experiments, and analyses.

2. LITERATURE REVIEW

In recent years, crowdsourcing has been widely applied to complete a variety of image labeling/classification tasks, from those requiring simple visual identification abilities to those that rely on domain expertise. Many studies have leveraged crowdsourcing to annotate large-scale datasets, often requiring subjective analysis such as conceptualized images (Nowak and Rüger, 2010), scene-centric images (Zhou et al., 2014), and general-purpose images from publicly available sources (Deng et al., 2009; Everingham et al., 2010). Crowdsourcing techniques have also been successfully tailored to many other complex visual labeling/classification contexts that require profound domain knowledge, including identifying fish and plants (He et al., 2013; Oosterman et al., 2014), endangered species through camera trap images (Swanson et al., 2015), locations of targets (Salek et al., 2013), land covers (Foody et al., 2018), and sidewalk accessibility (Hara et al., 2012). Due to its low cost and rapid processing capabilities, another prominent use of crowdsourcing is classification of CT images in medical applications. Such tasks have included identifying malaria-infected red blood cells (Mavandadi et al., 2012), detecting clinical features of glaucomatous optic neuropathy (Mitry et al., 2016), categorizing dermatological features (Cheplygina and Pluim, 2018), labeling protein expression (Irshad et al., 2017), and various other tasks (Nguyen et al., 2012; Mitry et al., 2013).

Despite its effectiveness at processing high work volumes, numerous technical challenges need to be addressed to maximize the benefits of the crowdsourcing paradigm. One such technical challenge involves deploying effective mechanisms for judgment/estimation aggregation, that is, the combining or fusing of multiple sources of potentially conflicting information into a single representative judgment. Since the quality of the predictions is highly dependent on the method employed to consolidate the crowdsourced inputs (Mao et al., 2013), a vast number of works have focused on developing effective algorithms to tackle this task. Computational social choice is a field dedicated to the rigorous analysis and design of such data aggregation mechanisms (Brandt et al., 2016). Researchers in this field have studied the properties of various voting rules, which have been applied extensively to develop better classification algorithms. The most commonly used method across various types of tasks is Majority Voting (MV) (Hastie and Kameda, 2005). MV attains high accuracy on simple idealized tasks, but its performance tends to degrade on those that require more expertise. One related shortcoming is that MV usually elicits and utilizes only one input from each participant—typically a binary response in crowdsourcing. Relying on a single form of input elicitation may decrease the quality of the collective judgment due to cognitive biases such as anchoring, bandwagon effect, decoy effect, etc. (Eickhoff, 2018). Studies have also found that the choice of input modality, for example, using rankings or ratings to specify a subjective response, can play a significant role in the accuracy of group decisions (Escobedo et al., 2022) and predictions (Rankin and Grube, 1980). These difficulties in data collection and aggregation mechanisms become even more prominent when the task at hand is complex (e.g., see Yoo et al., 2020). Researchers have suggested many potential ways of mitigating these limitations. One promising direction is the collection of richer data, i.e., using multiple forms of input elicitation. As a parallel line of inquiry, previous works suggest that specialized aggregation methods for integrating this data should be considered for making good use of these different pieces of information (Kemmer et al., 2020).

A logical enhancement of MV for the harder tasks is to elicit the participant's level of confidence (as a proxy of expertise) and to integrate these inputs within the aggregation mechanism. In the context of group decision-making, Grofman et al. (1983) suggested weighing each individual's inputs based on self-reported confidence of their respective responses, in accordance with the belief that individuals can estimate reliably the accuracy of their own judgments (Griffin and Tversky, 1992). More recently, Hamada et al. (2020) designed a wisdom of the crowds study that asked a set of participants to rank and rate 15 items they would need for survival and used weighted confidence values to aggregate their inputs. The results were sensitive to the size of the group (i.e., number of participants); when the group was small (fewer than 10 participants), the confidence values reportedly had little impact on the results. In a more realistic application, Saha Roy et al. (2021) used binary classification and stated confidence in these inputs to locate target objects in natural scene images. Their study showed that using the weighted average of confidence values improved collective judgment. It is important to remark that these and the vast majority of related studies incorporate the self-reported confidence inputs at face value. The Slating algorithm developed by Koriat (2012) represents a different approach that determines the response according to the most confident participant. For additional uses of confidence values to make decisions, we refer the reader to Mannes et al. (2014) and Litvinova et al. (2020).

Although subjective confidence values can be a valid predictor of accuracy in some cases (Matoulkova, 2017; Görzen et al., 2019), in many others they may degrade performance owing to cognitive biases that prevent a realistic assessment of one's abilities (Saab et al., 2019). Another natural approach is to weigh responses based on some form of worker reliability. Khattak and Salleb-Aouissi (2011) used trapping questions with expert-annotated labels to estimate the expertise level of workers. For domain-specific tasks where the majority can be systematically biased, Prelec et al. (2017) introduced the Surprisingly Popular Voting method, which elicits two responses from participants: their own answer and what they think the majority of other participants' answer is. It then selects the answer that is "more popular than people predict." Other aggregation approaches include reference-based scoring models (Xu and Bailey, 2012) and probabilistic inference-based iterative models (Ipeirotis et al., 2010; Karger et al., 2011).

In addition to crowdsourcing-based methods, automated image classification has become popular due to the breakthrough performances achieved by deep neural networks. Krizhevsky et al. (2012) used a convolutional neural network called AlexNet on a large dataset for the first time and achieved significant performance in image classification tasks compared to other contemporary methods. Since then, hundreds of studies have further improved classification capabilities, and a few have shown human-level performance when trained on large, noise-free datasets (Assiri, 2020; Dai et al., 2021). However, as the size and/or quality of training datasets decreases, the performance of these networks quickly degrades (Dodge and Karam, 2017; Geirhos et al., 2017).

A two-way relationship between AI and crowdsourcing can help compensate for some of the disadvantages associated with the two separate decision-making approaches. Human-elicited inputs interact with machine learning for a variety of reasons, but most are in service of the latter. A wider variety of ML models use human judgment to improve the accuracy and diversity in training data sets. For example, Chang et al. (2017) uses crowdsourcing to label images of cats and dogs since, unlike machines, humans can recognize these animals in many different contexts such as cartoons and advertisements. Human-elicited inputs are given more importance in specialized fields like law and medicine. For example, a study conducted by Gennatas et al. (2020) uses clinicians' inputs to improve ML training datasets and as a feedback mechanism using what is aptly termed "Expert-augmented machine learning." In a similarly promising direction, Hekler et al. (2019) uses a combination of responses from a user study and a convolutional neural network to classify images with skin cancer; the overall accuracy of their hybrid system was higher than both components in isolation.

Unlike human-AI interaction, human-AI collaboration is an emerging focus that can lead to the formulation of more efficient and inclusive solutions. Mora et al. (2020) designed an augmented reality shopping assistant that guides human clothing choices based on social media presence, historical purchase history, etc. As part of this focus, human-in-the-loop applications seek a more balanced integration of the abilities of humans and machines by sequentially alternating a feedback loop between them. For example, Koh et al. (2017) conducted a study where a field operator wearing smart glasses uses an artificial intelligence agent for remote assistance for hardware assembly tasks. Yet, few studies seek to combine human judgments and ML outputs to form a collective decision. Developing such equitable human-AI collaboration methods could be particularly beneficial in situations where the transparency, interpretability, and overall reliability of AI-aided decisions are of paramount concern.

3. CROWDSOURCING-BASED ML CLASSIFICATION

This section introduces different forms of input elicitations and describes how they can be utilized within a crowdsourcing-based ML classifier. Consider the image label aggregation problem where a set of images I are to be labeled by a set of participants P; without loss of generality, assume each image and participant has a unique identifier, that is, $I = \{i_1, i_2, ..., i_n\}$ and $P = \{p_1, p_2, ..., p_m\}$, where n and m represent the total number of images and participants, respectively. For each image $i_k \in I$, the objective is to infer the binary ground truth label $y_k \in \{0, 1\}$, where $y_k = 1$ if the specified target object is present in the image (i.e., positive image) and $y_k = 0$ otherwise (i.e., negative image). Since in these experiments each worker may label only a subset of the images, let $P(i_k) \subseteq P$ be the set of participants who complete the labeling task of image $i_k \in I$. In contrast to most crowdsourced labeling tasks where only a single label estimate is elicited per classification task, in the featured experiments each participant is asked to provide multiple inputs from the following five options. The first input is their binary response $l_k^j \in \{0, 1\}$ (i.e., classification label) indicating the presence/absence of the target object in image i_k. The second input is a coordinate-pair (u_k^j, v_k^j) indicating the location of the target object (elicited only when $l_k^j = 1$). The third input is a numeric value $c_k^j \in [0, 100]$ indicating the degree of confidence in the binary response l_k^j. The fourth input is another binary choice $g_k^j \in \{0, 1\}$ indicating what p_j estimates the binary response assigned by the majority of participants to i_k is; this input is referred to in this study as the Guess of Majority Elicitation (GME). The fifth input is a discrete rating $d_j^k \in \{1, 2, 3, 4\}$, whose values are mapped from four linguistic responses—1: "not at all difficult," 2: "somewhat difficult," 3: "very difficult," and 4: "extremely difficult"—indicating, in increasing order, the perceived difficulty of task i_k.

Before proceeding, it is worth motivating the use of participant confidence values in the proposed methods. Previous research has found that participants can accurately assess their individual confidence in their independently formed decisions (e.g., see Meyen et al., 2021). However, a pertinent concern regarding these confidence values is that, even if some participants are accurate in judging their performance at certain times, humans are generally prone to metacognitive biases, i.e., overconfidence or underconfidence in their actual abilities (Oyama et al., 2013). Hence, self-reported confidence should not be taken at face value, and specific confidence values should not be assumed to convey the same meaning across different individuals. In an attempt to mitigate such biases, the confidence values, $\{c_k^j\}_{k=1}^n$ provided by participant $p_j \in P$ are rescaled linearly between 0 and 100, with the lowest confidence value expressed by p_j being mapped to 0 and the greatest to 100. Letting $I^j \subseteq I$ be the set of images for which p_j provides a label, the confidence of participant p_j at classifying image i_k is rescaled as

$$c_k^{j*} = \frac{c_k^j - \min\limits_{i_q \in I^j} c_q^j}{\max\limits_{i_q \in I^j} c_q^j - \min\limits_{i_q \in I^j} c_q^j} \times 100.$$

The remainder of this section describes how the collected input elicitations are used as features in ML classifiers to generate predictions.

3.1. Features for Crowdsourcing-Based ML Methods

A total of seven features were extracted from the five inputs elicitations discussed in the beginning of this section for use with the ML classifiers; these features are described in the ensuing paragraphs.

- **Binary Choice Elicitation**: For each image $i_k \in I$, the binary choice elicitation values are divided into two sets: one containing the participants with response $l_k^j = 1$ and the other containing participants with response $l_k^j = 0$. The number of participants in each set can be used as an input feature within a ML classifier. However, since the number of participants can vary from image to image in practical settings, it is more prudent to use the relative size of the sets. Note that these relative sizes are complements of each other, that is, the fraction of participants who chose $l_k^j = 1$ as their binary choice label can be determined by subtracting from 1.0 the fraction of participants who chose $l_k^j = 0$. Therefore, to remove redundancy and co-linearity within the features, only one of these values is used as an input and is given as

$$x_k^1 = \frac{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(l_k^j = 1)}{|P(i_k)|},$$

where x_k^1 is the fraction of participants who specify that the target object is present in image i_k.

- **Spatial Elicitation**: A clustering-based approach is implemented to identify participants whose location coordinates (u_k^j, v_k^j)—elicited only when they specify that the target object is present—are close to each other. For each image $i_k \in I$, participants with binary choice label $l_k^j = 1$ are divided into multiple clusters using the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996). The reasons for choosing this algorithm are twofold. First, DBSCAN is able to identify groups of points that are close to each other but form arbitrary shapes; since the target images have varying shapes and sizes, this is what one would expect to see in a single image if all collected data points were overlaid onto it. Second, this clustering algorithm can easily mark as outliers/noise the points that are in low density areas, i.e., coordinate points that have significant distance from each other. After clustering, the fraction of participants belonging to the largest cluster is used as an input feature within the ML classifiers. For image i_k, this input feature can be expressed as

$$x_k^{SE} = \frac{\max\limits_{r \in R_k} n_r}{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(l_k^j = 1)},$$

where n_r is the number of participants in cluster r and R_k is the set of clusters identified by DBSCAN for image i_k.

- **Confidence Elicitation**: Although previous works have explored using confidence scores to improve annotation quality of crowdsourced data (Ipeirotis et al., 2010), very few have incorporated this input within a machine learning model. The confidence values are divided into two sets based on l_k^j, and the respective averages are used as additional features for the ML classifier. For image $i_k \in I$, these two input features can be expressed as

$$x_k^{conf,\,1} = \frac{\sum\limits_{p_j \in P(i_k)} c_k^{j*} \mathbb{1}(l_k^j = 1)}{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(l_k^j = 1)}; \qquad \text{and}$$

$$x_k^{conf,\,0} = \frac{\sum\limits_{p_j \in P(i_k)} c_k^{j*} \mathbb{1}(l_k^j = 0)}{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(l_k^j = 0)}.$$

Here, the confidence values are rescaled linearly between 0 and 100 before incorporating them as the features.

- **Guess of Majority Elicitation**: Similar to BCE, GME is converted into a single feature based on the number of participants whose g_k^j response value is 1 and is written as

$$x_k^{GME,\,1} = \frac{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(g_k^j = 1)}{|P(i_k)|}.$$

- **Perceived Difficulty Elicitation :** Previous research has shown that a task's perceived difficulty level can be used to some extent to improve the quality of annotation. In most cases, the difficulty level is set based on inputs from experts, that is, participants with specialized knowledge with respect to the task at hand (Khattak and Salleb-Aouissi, 2011), or it is estimated from the classification labels collected from participants (Karger et al., 2011). Unlike these works, the featured experiments gather the perceived difficulty of each task directly from each participant to evaluate the reliability of this information and its potential use within ML classifiers. For each image $i_k \in I$, the difficulty elicitation values d_k^j are divided into two sets: one for the participants with response $l_k^j = 1$, and

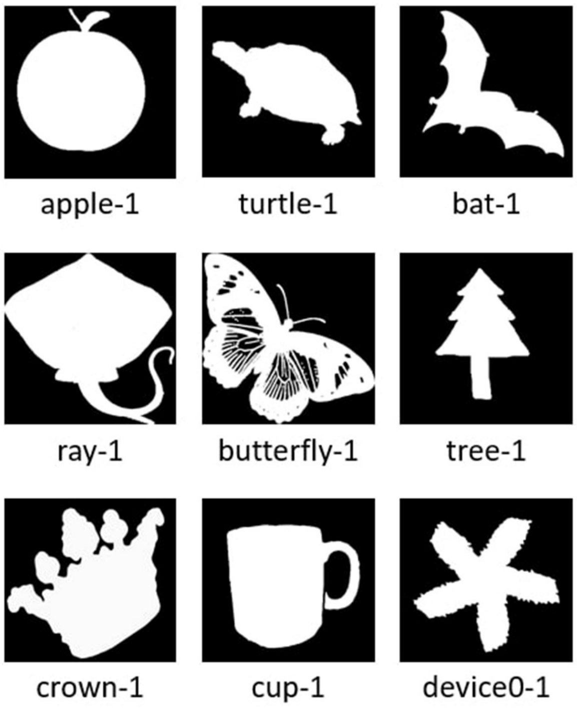


FIGURE 1 | Object/shape templates from the MPEG-7 core experiment CE-shape-1 test set.

the other for the remaining participants with response $l_k^j = 0$. The average values from each set are then used as additional features for the ML classifier; these two input features can be expressed as

$$x_k^{PDE,\,1} = \frac{\sum\limits_{p_j \in P(i_k)} d_k^j \mathbb{1}(l_k^j = 1)}{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(l_k^j = 1)}; \quad \text{and}$$

$$x_k^{PDE,\,0} = \frac{\sum\limits_{p_j \in P(i_k)} d_k^j \mathbb{1}(l_k^j = 0)}{\sum\limits_{p_j \in P(i_k)} \mathbb{1}(l_k^j = 0)}.$$

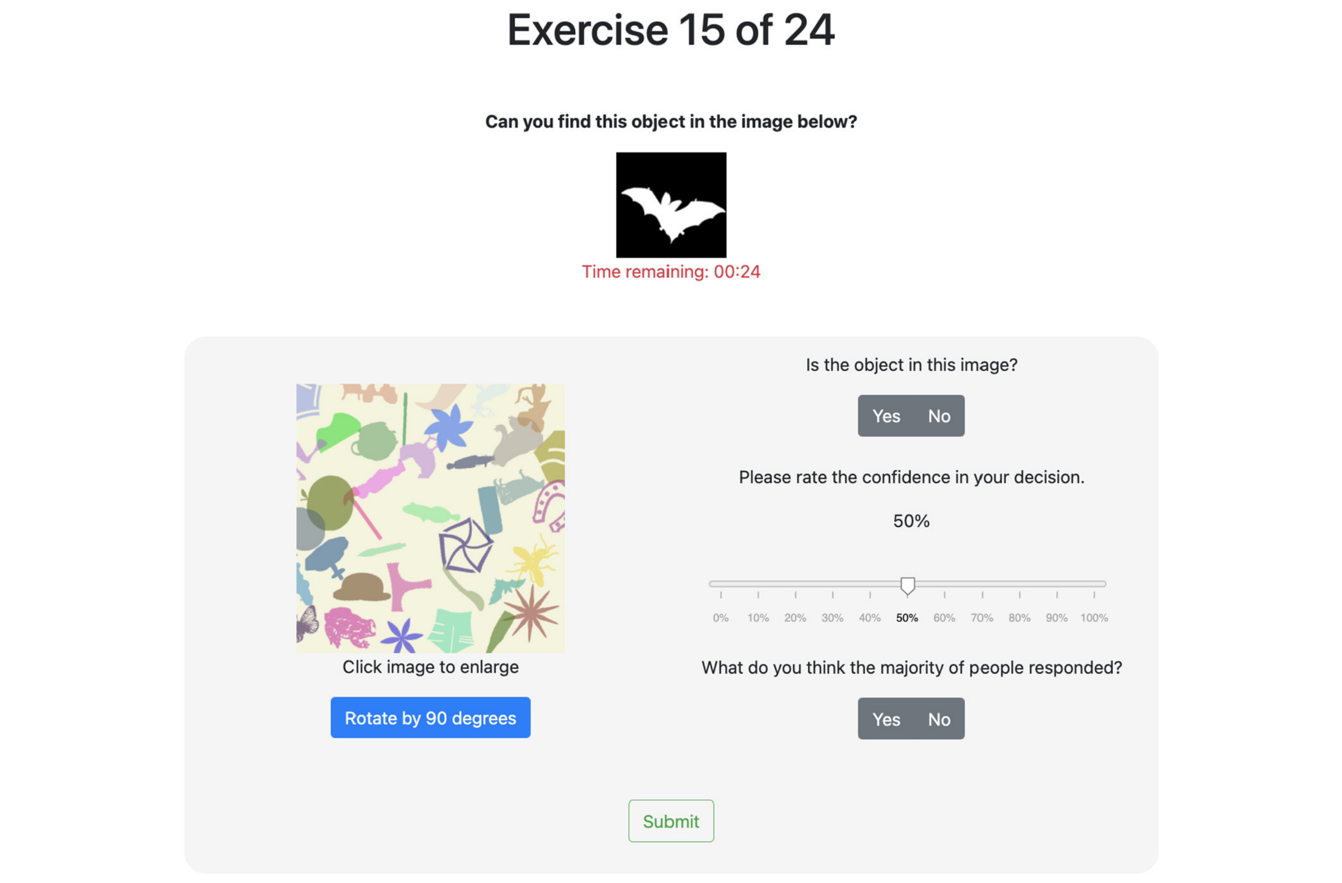
4. EXPERIMENT DESIGN

Prior to introducing the components of the experiment design, we describe the *MPEG-7 Core Experiment CE-Shape-1 Test Set* (Jeannin and Bober, 1999; Ralph, 1999), which is the source data from which the featured crowdsourcing activities are constructed. The dataset is composed of black and white images of a diverse set of shapes and objects including animals, geometric shapes, common household objects, etc. In total, the dataset consists of 1, 200 objects/shapes (referred to here as *templates*) divided into 60 object/shape classes, with each class containing 20 members. **Figure 1** provides representative templates from some of these classes.

The images used in the crowdsourcing experiment are constructed by instantiating and placing multiple MPEG-7 Core Experiment CE-Shape-1 Test Set templates onto a single image frame. The instantiation of the image template is specified with six adjustable parameters: density, scale, color, transparency, rotation, and target object. See **Supplementary Material** for a detailed description of these parameters.

4.1. Description of Activities

For the crowdsourcing activities, we designed two studies, each of which elicits multiple forms of input from participants to complete a number of image classification tasks. A user interface was designed and implemented to perform the two studies, which differ based on the subsets of input elicitations tested and the class balance ratios of the image datasets (more details are provided later in this subsection). The interfaces were developed in HTML and Javascript and then deployed using Amazon Mechanical


Exercise 15 of 24

Can you find this object in the image below?

Time remaining: 00:24

Is the object in this image?

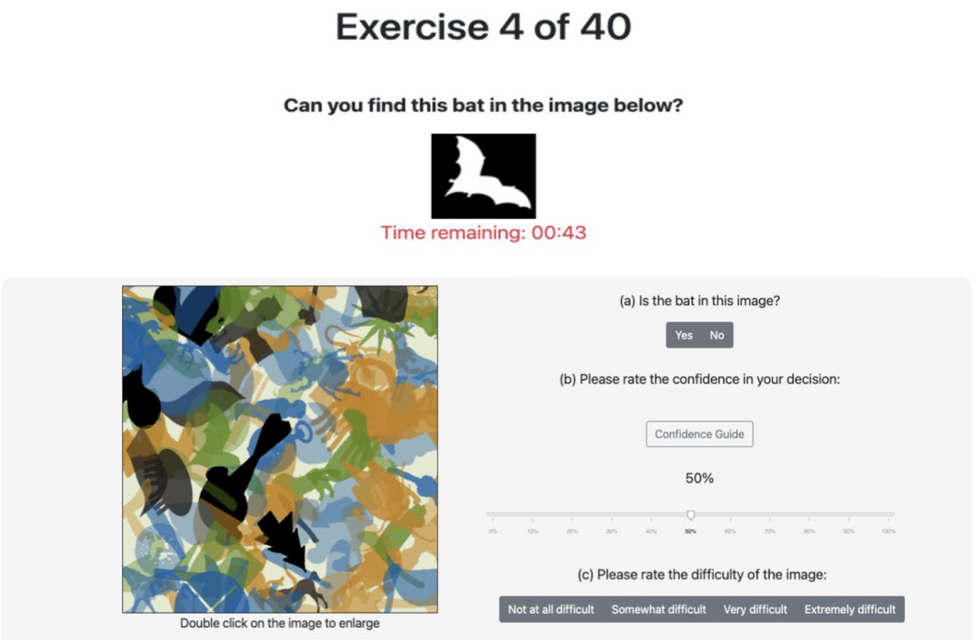
Please rate the confidence in your decision.

50%

Click image to enlarge

What do you think the majority of people responded?

FIGURE 2 | Image classification task UI for balanced dataset—image contains bat (lower right).


Exercise 4 of 40

Can you find this bat in the image below?

Time remaining: 00:43

(a) Is the bat in this image?

(b) Please rate the confidence in your decision:

50%

(c) Please rate the difficulty of the image:

Double click on the image to enlarge

FIGURE 3 | Image classification task UI for imbalanced dataset—image contains bat (center left).

TABLE 1 | Summary of experiment image parameters.

Exp.		Images	Density	Scale	Color	Transparency	Target
Set A	#1	16	{100, 120, 140, 160}	$\{T(0.2 \pm 0.12), .., T(0.65 \pm 0.12)\}$	Discrete: {4}	$U(100, 200)$	Bat
	#2						Butterfly
	#3						Apple
	#4						Stingray
Set B	#5	24	{80}	$\{T(0.2 \pm 0.05), T(0.3 \pm 0.05)\}$	Discrete: {1,...,6}	$U(140, 170)$	Bat
	#6		{80,100,120}	$\{T(0.2 \pm 0.05), .., T(0.4 \pm 0.05)\}$	$U(10, 255)$ for R,G,& B		Turtle
	#7		{100, 150}	$\{T(0.2 \pm 0.05), T(0.3 \pm 0.05)\}$			Various-7
Set C	#8	40	{90, 100, 115, 150}	$\{T(0.25, 0.35, 0.40)\}$	Discrete: {4}	$U(150, 200)$	Bat
	#9						
	#10						
Set D	#11						
	#12						
	#13						

Turk (MTurk). Participants were first briefed about the nature of the study and shown a short walk-through video explaining the interface. Afterwards, participants proceeded to the image classification tasks, which were shown in a randomized order. After completing an experiment, participants were disallowed to participate in further experiments. **Figures 2**, **3** provide examples of the user interfaces, both of which instituted a 60 s time limit to view each image before it was hidden. If the participant completed the input elicitations before the time limit, they were allowed to proceed to the next image; on the other hand, if the time limit was reached, the image was hidden from view but participants could take as much time as they needed to finish providing their inputs. The time limit was imposed to ensure the scalable implementation of a high number of tasks. In particular, the goal is to develop activities that can capture enough quality inputs from participants while mitigating potential cognitive fatigue. In preliminary experiments, we found that participants rarely exceeded 45 s. In the featured studies (to be described in the next two paragraphs), the full 60 s were utilized in only 7% of the tasks, with an average time of around 27 s. The number of tasks given to the participants varied by experiment and ranged from 16 to 40 images (see **Table 1** for details). We deemed this number of tasks to be reasonable and not cognitively burdensome to participants based on findings of prior studies with shared characteristics. For instance, Zhou et al. (2018) performed a visual identification crowdsourcing study where participants were assigned up to 80 tasks, each of which took a median time of 29.4 s to complete. The authors found that accuracy decreased negligibly for this workload (i.e., twice as large as in the featured studies).

In the first study, seven experiments were completed and grouped into two sets: Experiment Set A (four experiments) and Experiment Set B (three experiments). Each experiment used a balanced set of images, with half containing the target template (i.e., positive images); target objects were chosen so as to avoid confusion with other template classes. See **Table 1** for image generation parameters, and see **Supplementary Material** for additional related details. The parameter ranges selected for Experiment Set A were designed to keep the difficulty of the classification tasks relatively moderate. On the other hand, a more complex set of parameters was selected for Experiment Set B to expand the range of difficulty. These differences are reflected in the individual performance achieved in these two experiment sets, measured by the respective average number of correct classifications obtained by participants. For Experiment Set A, individual performance averages ranged between 59 and 77% for each of the four experiments, whereas for Experiment Set B, they were between 54 and 82% for each of the three experiments.

In the second study, six experiments were conducted. These experiments were also grouped into two sets: Experiment Set C (three experiments) and Experiment Set D (three experiments). Each consisted of image sets with an imbalanced ratio of positive-to-negative images. Experiment Set C had a 20-80 balance, meaning that 20% of the images were positive, and 80% were negative; Experiment Set D had a 10–90 balance. The results of Experiment Sets A and B revealed that *scale* and *density* are the only factors that had a statistically significant impact on individual performance. Based on this insight, we constructed a simple linear regression model with these two parameters as the predictors and *proportion of correct participants* as the responses; the model is very significant ($p < 0.001$), and its adjusted R-squared value is 0.65. The model was used to generate image sets with an approximated difficulty level by modifying the scale and density parameters accordingly. It should be noted that the true difficulty of each image varies based on the random generation process. The model was implemented to design experiments consisting of classification tasks of reasonable difficulty—that is, neither trivial nor impossible to complete. Images of four levels of difficulty were generated for Experiment Sets C and D. At each difficulty level, the density was varied while keeping the other parameters consistent across images. This resulted in images that appear similar, but with different amounts of "clutter". The four difficulties generated were categorized as "very difficult," "difficult," "average," and "easy." See **Supplementary Material** for details and sample images of each difficulty. Experiment Sets C and D use an even split of each difficulty (i.e., 25% of generated images from each level). For the three respective experiments, individual average accuracy values ranged between 65 and 73% for Set C and between 58 and 72% for Set D.

Figures 2, **3** show the user interface presented to participants in the first and second study, respectively. For each classification task (i.e., image) in the first study, participants were asked to provide a binary response indicating whether or not a target object is present. If they responded affirmatively, they were then prompted to locate the target object by clicking on it. Then, participants were asked to rate their confidence in their binary response on a scale from 0 to 100%. Finally, participants were asked to guess the binary response of the majority of participants. The second study asked participants similar questions as the first study. For each classification task, participants were also asked to provide a binary responses indicating whether or not a target object is present and their level of confidence in this response. If they responded affirmatively, however, they were then prompted to locate the target object by drawing a bounding box around it; the centroid of the bounding box was used as the (x, y)-coordinate gathered from this elicitation. In replacement to the last question of the first study, participants were asked to rate the difficulty of the specific image being classified based on a discrete scale. The rating choices provided were "not difficult at all," "somewhat difficult," "very difficult," and "extremely difficult." These labels were mapped to 1, 2, 3, and 4, respectively, for use in the aggregation algorithms.

4.2. Participant Demographics and Filtering of Insincere Participants

A total of 356 participants were recruited and compensated for their participation using Amazon MTurk. Participants in Experiment Set A were paid $1.25, those in Experiment Set B were paid $2.00, and those in Experiment Sets C and D were paid $3.75. The differences in compensation can be attributed to the number of questions and the difficulty of image classification tasks of the respective experiment sets. Participants were made aware of the compensation amount before beginning the study. Payment was based only on completion and not on performance. Before proceeding, it is necessary to delve further into the quality of the participants recruited *via* the MTurk platform, and the quality of data they provide. Because of the endemic presence in most crowdsourcing platforms of annotators who do not demonstrate an earnest effort (Christoforou et al., 2021), some criteria should be defined to detect such insincere participants and filter out low-quality inputs. This work defined two criteria for characterizing (and filtering out) an annotator as insincere:

- **Criterion 1:** The participant answered over 75% of the questions in no more than 10 s per question.
- **Criterion 2:** The participant's binary responses were exclusively 0 or exclusively 1 over the entire question set.

Criteria 1 was imposed based on the following reasoning. In general, classification of negative images takes longer than classification of positive images. Even if it is assumed that participants can spot the positive images immediately (i.e., within 10 s), it should take more than 10 s to reply to the negative images that are of moderate to high difficulty. Because each Experiment Set in this study contained at least 50% negative images (Experiment Sets C and D contain a higher percentage) and only a small minority were of low difficulty, a conservative estimate that participants should take longer than 10 s to answer

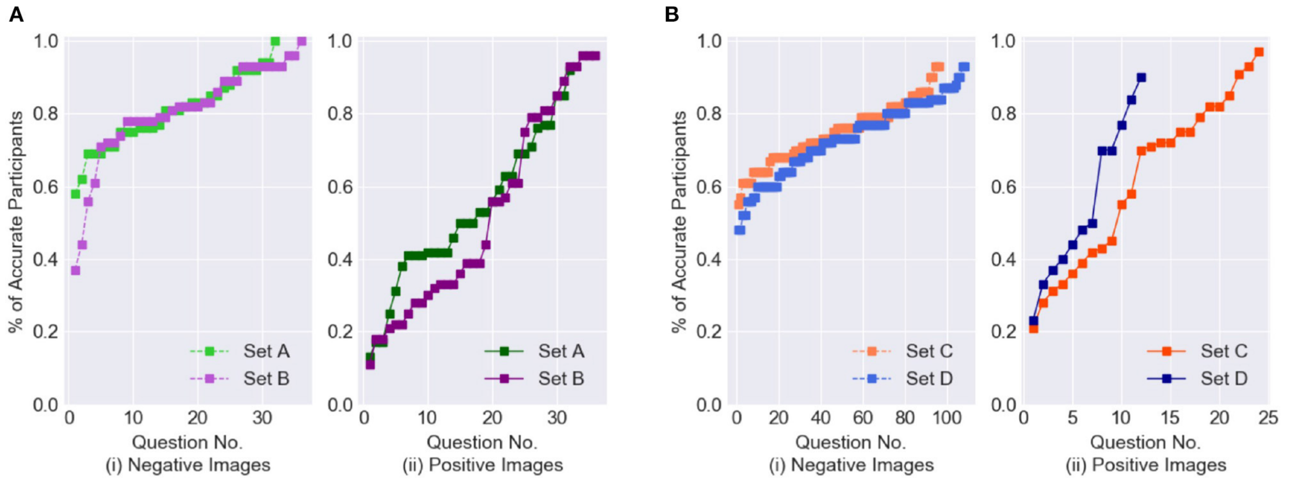


FIGURE 4 | Distribution of Binary Classification results from crowdsoured data. **(A)** Balanced dataset. **(B)** Imbalanced dataset.

at least 25% of the images was set (i.e., to be more lenient toward the participants). Further analysis of the behavior of the participants in relation to the task completion times supporting this observation has been added to **Supplementary Material**.

From the initial 356 participants, 50 participants were removed from the four experiment sets using the above criteria. Among them, 15 fell under criterion 1 and the rest under criterion 2. As expected, filtering out these data provided less noisy inputs to the crowdsourcing-based aggregation methods. From the remaining 306 participants, 276 completed the demographics survey. Their reported ages ranged from 21 to 71 years old, with a mean and median of 36 and 33, respectively. 156 participants reported their gender as male, 120 as female, and 0 as other. In terms of reported education level, 23 participants finished a high school/GED, 17 some college, 16 a 2-year degree, 148 a 4-year degree, 70 a master's degree, 1 a professional degree, and 2 a doctoral degree.

4.3. Distribution of Crowdsourced Data

Before proceeding to the computational results, it is pertinent to analyze the data collected from the crowdsourcing experiments. First, let us analyze the relationship between the perceived difficulty levels reported by the participants (i.e., input feature PDE) and the difficulty levels utilized in the proposed image generation algorithm (see Section 4.1 for details). The average difficulty values reported by participants for images categorized by the algorithm as "very difficult," "difficult," "average," and "easy" were 2.89, 2.73, 2.62, and 2.03, respectively. This evinces a clear correlation, with the "very difficult" images having the highest average perceived difficulty values and the rest reflecting a decreasing order of difficulty, which supports the ability of the image generation method used in this study to control the classification task difficulty, according to the four above-mentioned categories.

Next, let us analyze the correctness of the binary response values collected from the participants. **Figure 4** shows the percentage of participants who answered each question accurately; question numbers have been reordered for each of the four datasets by increasing participant accuracy. The positive and negative images for the balanced and imbalanced datasets are presented in separate graphs. The plots show that, for the balanced datasets (Experiments Sets A and B), the accuracy on the positive images is significantly lower than on the negative images. Moreover, in Experiment Set B, nearly half of the positive images have accuracy values below 0.4, whereas in Experiment Set A most images have values above 0.4. This is a good indication of the higher difficulty level of Experiment Set B. For the imbalanced datasets, in both Experiments Sets C and D, nearly all negative images have accuracy values above 0.4. In Experiment Set C, there is an almost even distribution of the positive images above and below 0.6, whereas in Experiment Set D nearly 60% of the positive images have accuracy values below 0.5, indicating that Experiment Set D was comparatively more difficult.

5. COMPUTATIONAL RESULTS

This section compares the performance of the voting and crowdsourcing-based ML methods presented in Section 3 on both balanced and imbalanced datasets. As a baseline of comparison for the proposed crowdsourcing-based ML methods, three traditional voting methods are used: Majority Voting (MV), Confidence Weighted Majority Voting (CWMV), and Surprisingly Popular Voting (SPV). The details of these methods can be found in **Supplementary Material**. For the ML methods, four binary classification approaches were selected: K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Forest Classifier (RF), and Linear Support Vector Machines

(SVM-Linear). These were selected as reasonable representatives of commonly available methods. The ML classifiers were trained and evaluated using built-in functions of the Python *scikit-learn* library (Pedregosa et al., 2011). The hyper-parameters were optimized on a linear grid search with a nested 5-fold cross-validation strategy. However, due to the small size of the datasets, a Leave-One-Out (LOO) cross-validation strategy was used to train and evaluate the classifiers.

In the DBSCAN clustering approach used for extracting the Spatial Elicitation (SE), the maximum distance between two data points in the cluster (ϵ) and the minimum data points required to form a cluster (*MinPts*) was set to 50 and 3, respectively. The former was set based on the size of the target objects used relative to the size of the image frame ($1,080 \times 1,080$); the latter was set to ensure a sufficiently low probability of forming a cluster with random inputs. To obtain a rough estimate of this probability, consider the case where three participants with binary response $l_k^j = 1$ randomly select their location coordinates on an image with area A. The probability of two points having a maximum distance of r (i.e., falling within a circle with radius r) is $\pi r^2/A$ and, therefore, the probability of the three points being identified as a cluster by DBSCAN is $2(\pi r^2/A)^2$. Setting $r = \epsilon = 50$ and $A = 1,080 \times 1,080$ for our experiment, this probability value becomes 0.01, which is sufficiently small and justifies the use of the selected parameters.

TABLE 2 | Performance analysis of voting methods for balanced dataset.

	MV		CWMV		SPV	
	Acc.	FNR	Acc.	FNR	Acc.	FNR
Experiment Set A	0.73	0.53	0.81	0.34	0.45	0.94
Experiment Set B	0.71	0.53	0.74	0.47	0.53	0.92

5.1. Performance of Aggregation Methods on Balanced Datasets

This section compares the performance of the voting and ML methods on balanced datasets (Experiment Sets A and B). The initial study elicits four out of the five inputs listed in Section 3.1: BCE, GME, CE, and SE. The results are summarized in **Tables 2** and **3**.

The performance of the ML methods is quantified *via* three performance metrics: accuracy (Acc.), false-negative rate (FNR), and area under the ROC curve (AUC). For the voting methods, only the first two of these metrics are reported. For each of the ML classifiers, the best accuracy, FNR, and AUC values among the different input elicitation combinations are marked in bold. Before proceeding, it is worthwhile to mention two additional points regarding the values presented in the table. First, each row in **Table 3** represents a different combination of inputs used as features for the ML classifiers. For example, BCE-CE indicates that both binary and confidence elicitation inputs (i.e., $x_k^1, x_k^{conf,\,1}$ and $x_k^{conf,\,0}$) were used as features for the ML classifiers, whereas BCE-CE-SE-GME indicates that all four input elicitations (i.e., $x_k^1, x_k^{conf,\,1}, x_k^{conf,\,0}, x_k^{SE}$, and $x_k^{GME,\,1}$) of Experiment Set A and B were used as the respective input features. Second, when calculating the accuracy and FNR values of the voting methods, images with undecided outcomes (i.e., ties) are considered as a third separate label.

Let us first discuss the performance of the aggregation models in terms of accuracy. For Experiment Sets A and

TABLE 3 | Performance analysis of crowdsourcing based ML methods for balanced dataset.

Input	KNN			LR			RF			SVM-Linear		
Elicitations	Acc.	FNR	AUC	Acc.	FNR	AUC	Acc.	FNR	AUC	Acc.	FNR	AUC
					Experiment Set A							
BCE	0.83	**0.16**	0.87	**0.89**	**0.16**	**0.95**	0.83	**0.19**	0.86	**0.89**	**0.16**	0.90
BCE-CE	**0.86**	0.22	**0.89**	0.86	0.19	0.89	**0.84**	0.22	**0.93**	0.86	0.19	0.91
BCE-SE	0.84	0.22	0.85	0.88	**0.16**	0.91	0.83	0.22	0.87	0.86	0.19	0.92
BCE-GME	**0.86**	0.19	0.87	0.86	**0.16**	0.91	0.83	**0.19**	0.83	0.88	**0.16**	0.91
BCE-CE-SE	0.81	0.31	0.86	0.88	0.19	0.88	**0.84**	0.22	0.91	**0.89**	**0.16**	0.91
BCE-CE-GME	0.8	0.25	0.82	0.84	0.19	0.90	0.83	0.22	0.89	0.84	0.19	0.90
BCE-CE-SE-GME	**0.86**	0.25	0.82	0.83	0.19	0.90	0.83	0.22	0.89	0.86	0.19	0.89
					Experiment Set B							
BCE	0.75	0.28	0.79	0.81	0.28	0.74	0.75	0.31	0.76	0.74	0.42	0.85
BCE-CE	0.78	0.28	**0.85**	0.81	0.25	0.88	0.75	**0.22**	**0.82**	**0.79**	**0.25**	0.85
BCE-SE	**0.79**	**0.19**	0.81	0.68	0.42	0.55	0.74	0.31	0.78	0.74	0.44	0.80
BCE-GME	0.75	0.31	0.78	0.76	0.31	**0.89**	0.68	0.33	0.74	0.72	0.42	**0.88**
BCE-CE-SE	0.76	0.22	0.79	**0.82**	**0.22**	**0.89**	0.74	0.25	0.80	0.72	0.47	0.85
BCE-CE-GME	0.76	0.31	0.81	0.78	0.28	0.80	**0.76**	**0.22**	0.79	0.78	0.31	0.86
BCE-CE-SE-GME	0.72	0.36	0.82	0.78	0.31	0.87	0.72	0.31	0.79	0.74	0.47	0.83

Bold values denote best performance among the different input elicitation combinations for each Crowdsourcing-based ML method.

B, the average accuracy value of MV was stable at around 72%. The CWMV method performed significantly better than MV, achieving an average accuracy value of around 77%. SPV was the worst performer across the board, with an average accuracy value of <50% (i.e., worse than a purely random classifier). This low performance can be largely attributed to the excessive number of tied labels generated compared to the other methods. In SPV, 18 out of the 136 instances were classified as tied (i.e., participants were undecided regarding the guess of the majority's estimate). By comparison, there were only three tied instances with MV and none with CWMV.

The results of the ML classifiers in Experiment Set A were relatively consistent in terms of both accuracy and AUC values for all seven combinations of the input elicitations. The classifiers performed particularly well, attaining accuracy values above 83% for all combinations; this can be partly explained by the fact that the images in this experiment set were generated using parameter ranges that were more consistent and less variable in difficulty. In Experiment Set B, the ML classifiers reached higher accuracy and AUC values under certain combinations of the input elicitations. For RF, LR, and KNN, a noticeable increase in AUC values (from 76 to 85%) results when using the BCE-CE combination compared to the standalone BCE input; the accuracy values in these cases either increased or stayed the same. Altogether, these results suggest that integrating CE into an ML classifier can help attain more accurate predictions when the sample size is small and the difficulty level of the images is more varied. Furthermore, they show that the ML classifiers outperformed the voting methods, with the LR classifier achieving the highest values in terms of both accuracy and AUC scores.

Another performance metric of interest is FNR, which denotes the fraction of images the methods label as 0 (i.e., negative) when their true label is 1 (i.e., positive). A high FNR may be concerning in many critical engineering and medical applications where a false-negative may be more detrimental than a false-positive since the latter can be easily verified in subsequent steps. For example, FNR has significant importance in detecting lung cancer from chest X-rays. If the model falsely classifies an X-ray as negative, the patient may not receive needed medical care in a timely fashion. Returning to **Table 2**, the FNRs of the three voting methods are high across the board, with SPV again having the worst performance. The high FNRs of MV and CWMV can be attributed to the fact that people tend to label the image as negative whenever they fail to find the target object and that these methods are unable to extract additional useful information from the responses.

In Experiment Set A, the accuracy values are the highest for the BCE-CE combination, whereas the FNR values are the lowest for the single BCE input. On the other hand, in Experiment Set B, although the accuracy values are the same for both input combinations, FNR values decrease for the BCE-CE combination. Moreover, for SVM, the reduction in FNR values is significant for Experiment Set B (from 42 to 25%) for the BCE-CE combination. This outcome reiterates the advantages of integrating CE into ML classifiers for more complex datasets.

5.2. Performance of Aggregation Methods on Imbalanced Datasets

This section compares the performance of the voting and crowdsourcing-based ML methods on imbalanced datasets (Experiment Sets C and D). Similar to the balanced datasets, a total of four input elicitations are utilized. However, for this study, the GME input is replaced by the PDE input (i.e., a rating value to assess the difficulty of the classification task), which is explained as follows. Recall from the discussion of Section 5.1 that none of the ML classifiers obtained a performance improvement when using the GME input relative to the other elicitation combinations. Moreover, the only method that utilizes the GME elicitation, SPV, was the worst-performing among the three voting methods. The inability of the GME input to provide any additional information during the classification process prompted its removal from subsequent studies. Due to this modification, only two voting methods (MV and CWMV) are explored for the imbalanced datasets.

When the dataset is balanced, accuracy by itself is a good indicator of the model's performance. However, when the dataset is imbalanced, accuracy can often be misleading as it provides an overly optimistic estimation of the classifier's performance on the majority class ("0" in this experiment). In such cases, a more accurate evaluation metric is the F_1-score (Sokolova et al., 2006), defined as the harmonic mean of the precision and recall values and can be expressed as, $F_1\text{-score} = 2(\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) = \text{TP}/[\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})]$, where, TP, FP, and FN refers to the number of true-positives (images the methods label as 1 when their true label is 1), false-positives (images the methods label as 1 when their true label is 0), and false-negatives (images the methods label as 0 when their true label is 1), respectively. Since both Experiment Sets C and D are highly imbalanced (with an average of 15% of their images belonging to the positive class) the F_1-score is reported instead of accuracy to better estimate the performance of the classifiers.

The overall results for the voting and machine learning methods are summarized in **Tables 4**, **5**, respectively. The performance of the ML methods is quantified *via* three performance metrics: F_1-score, FNR, and AUC; for the voting methods, only the first two of these metrics are reported. Let us first discuss the performance of the aggregation methods in terms of F_1-score. For Experiment Sets C and D, MV and CWMV have comparable scores, with both having the same value in the first set and MV outperforming CWMV by a slight margin in the second set. Moving on to the ML methods, for Experiment Set C, the ML classifiers displayed comparable F_1-scores for

TABLE 4 | Performance analysis of voting methods for imbalanced dataset.

	MV		CWMV	
	F_1	FNR	F_1	FNR
Experiment Set C	0.77	0.38	0.77	0.25
Experiment Set D	0.53	0.58	0.52	0.50

TABLE 5 | Performance analysis of crowdsourcing based ML methods for imbalanced datasets.

Input Elicitations	KNN			LR			RF			SVM-Linear		
	F_1	FNR	AUC	F_1	FNR	AUC	F_1	FNR	AUC	F_1	FNR	AUC
						Experiment Set C						
BCE	0.73	0.38	0.82	0.78	0.33	0.79	0.73	0.33	0.81	0.73	0.38	0.92
BCE-CE	0.75	0.38	0.89	0.81	0.29	0.90	0.78	0.33	0.86	0.80	0.33	0.90
BCE-SE	**0.81**	**0.29**	0.83	**0.84**	**0.25**	**0.95**	0.76	**0.29**	0.87	**0.84**	**0.25**	0.86
BCE-PDE	**0.81**	**0.29**	0.83	0.76	0.33	0.94	0.68	0.38	0.81	0.77	0.38	0.9
BCE-CE-SE	0.76	0.33	**0.92**	0.81	0.29	0.92	0.77	**0.29**	**0.90**	0.81	0.29	0.88
BCE-CE-PDE	**0.81**	**0.29**	0.90	0.81	0.29	0.86	0.79	**0.29**	0.86	0.80	0.33	0.90
BCE-CE-SE-PDE	**0.81**	**0.29**	**0.92**	0.81	0.29	0.86	**0.81**	**0.29**	**0.90**	0.81	0.29	0.86
						Experiment Set D						
BCE	0.53	**0.58**	0.59	0.55	**0.33**	**0.87**	0.36	0.58	0.64	0.61	**0.42**	0.85
BCE-CE	**0.59**	**0.58**	**0.76**	0.54	0.42	0.83	**0.63**	**0.50**	**0.79**	**0.67**	0.50	**0.87**
BCE-SE	**0.59**	**0.58**	0.62	0.46	0.50	0.84	0.36	0.58	0.65	0.63	0.50	0.8
BCE-PDE	0.50	0.67	0.67	**0.57**	**0.33**	0.85	0.47	0.67	0.78	0.56	**0.42**	0.86
BCE-CE-SE	**0.59**	**0.58**	0.72	0.52	0.42	**0.87**	0.53	0.58	0.77	**0.67**	0.50	0.84
BCE-CE-PDE	0.44	0.67	0.68	0.56	0.42	0.73	0.53	0.58	**0.79**	0.63	0.50	**0.87**
BCE-CE-SE-PDE	0.56	**0.58**	0.74	0.52	0.42	0.84	0.44	0.67	0.78	**0.67**	0.50	0.85

Bold values denote best performance among the different input elicitation combinations for each Crowdsourcing-based ML method.

combinations BCE-CE, BCE-SE, BCE-CE-SE, and BCE-CE-SE-PDE. In addition, all four of these input combinations performed better than the standalone BCE input. The RF and KNN classifiers achieved the highest values with the combination BCE-CE-SE-PDE. In contrast, the LR and SVM classifiers achieved the highest values with the BCE-SE combination. Overall, the LR classifier achieved the best performance for this set with inputs BCE-SE. In Experiment Set D, the results followed a different pattern. In this case, the classifiers achieved the same or higher values when the BCE-CE combination was used compared to the BCE-SE or BCE-CE-SE combinations, indicating that the SE input does not provide any additional information for this experiment set. Because this dataset is highly skewed toward the negative class (10–90 balance), we conjecture that participants may have become demotivated to closely inspect difficult images from the positive class. Whatever the cause, smaller clusters were obtained from these images, reducing the effectiveness of the SE input in many cases. In Experiment Set D, the highest performance was achieved by the SVM classifier for the BCE-CE input. These results once again indicate that, even though the self-reported confidence values are not particularly helpful when used within the traditional voting methods context (Li and Varshney, 2017; Saab et al., 2019)—as can also be seen by the performance of the CWMV algorithm in this study—incorporating them into an ML classifier can help attain better performance, specifically higher F_1-scores for highly imbalanced datasets.

In terms of FNR, the performance of the CWMV method was markedly better than the MV method for both Experiment Sets. The assigned labels for the positive images in Experiment Set D for the two voting methods are almost identical, with the exception of a single image which the latter labeled as a tie (i.e., undecided), contributing to the decrease in performance. Note that none of the images in Experiment Set C was labeled as a tie by either of the voting methods. Among the ML methods, LR significantly outperformed all of the other classifiers for Experiment Set D. Although in Experiment Set C the FNR for the BCE-SE combination (25%) was lower than for the BCE combination (33%), in Experiment Set D a significant increase (33–50%) can be seen between these two combinations. Overall, ML classifiers outperformed MV; however, CWMV showed comparable performance for both experiment sets. Note that a distinctive advantage of CWMV over the ML methods is that it does not require training data.

5.3. Changing the Threshold of Positive Classification

This section examines how voting methods can be modified to emphasize other important metrics of image classification. In particular, it seeks to prioritize reduced false-negative rates, which are relevant in various critical applications. The FNRs can be reduced by lowering the threshold at which a positive classification is returned by a classification method (i.e., changing the tipping point for returning a positive collective response). However, care must be exercised when lowering the threshold since this implicitly increases false-positive rates (FPRs), which can also be problematic.

By default, the threshold at which voting methods return a positive response is fixed; for example, MV requires more than 50% of positive responses to return the positive class. **Figure 5** illustrates the impacts of adjusting the thresholds for the voting methods as well as for the ML methods; the figure separates FNRs from FPRs for each method. Using MV as an example, decreasing the threshold from 0.5 to 0.3 results in relatively small increases to the FPR and larger decreases to the FNR; further decreases cause a disproportionate increase to FPRs. Hence, these inflection points can help guide how the thresholds can be set for each

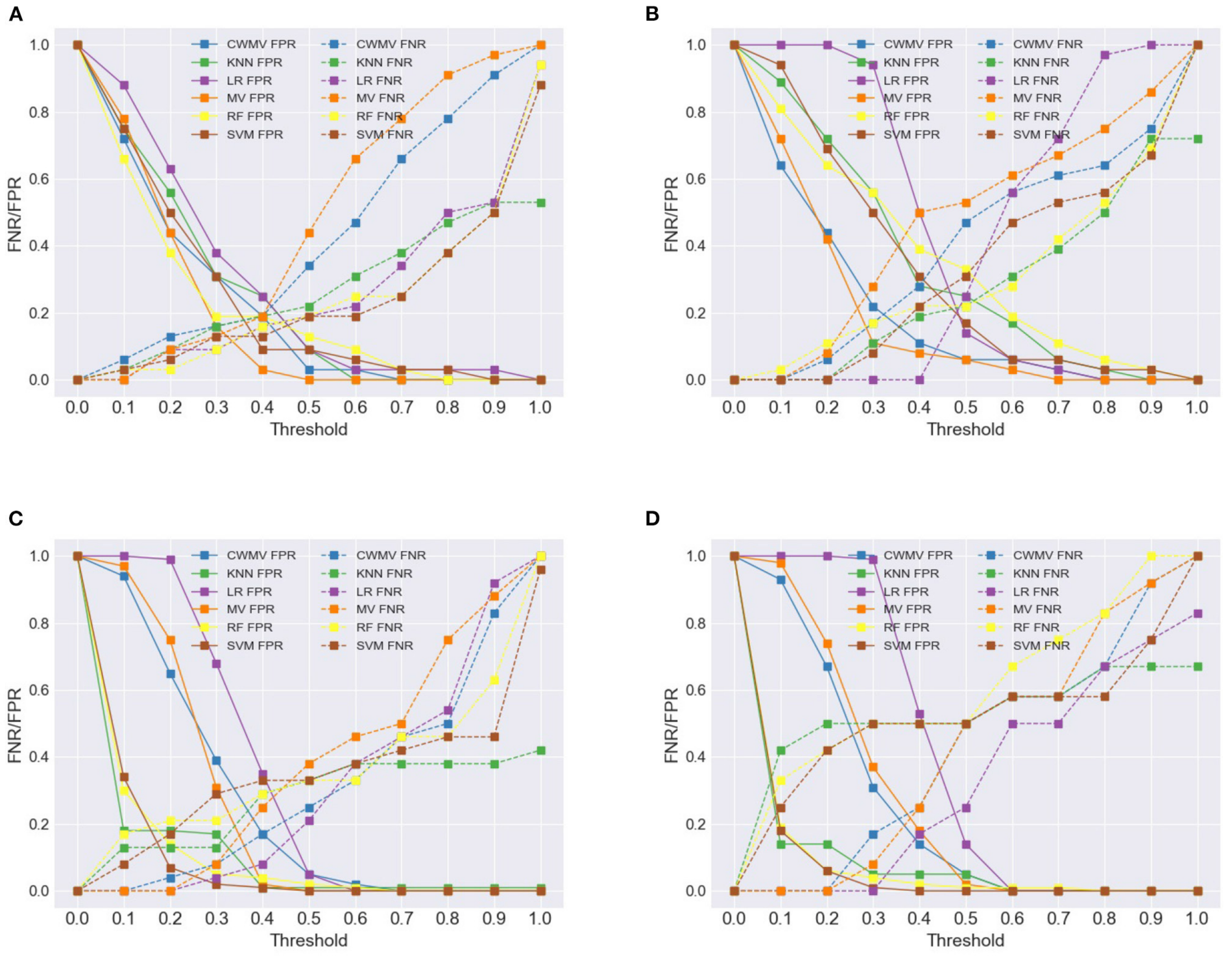


FIGURE 5 | Change in FNR/FPR of different aggregation methods under varying thresholds. **(A)** Experiment Set A, **(B)** Experiment Set B, **(C)** Experiment Set C, **(D)** Experiment Set D.

voting method to prioritize FNR. A similar observation can be made about the FNRs of the ML methods (except for LR) for the imbalanced datasets. However, this does not hold for the ML methods for the balanced datasets—for example, reducing the threshold to 0.3 causes a significant increase in FPRs compared to the decrease in FNRs. This suggests that caution must be exercised when changing the threshold of positive classification of ML classifiers.

6. ENHANCEMENT OF CROWDSOURCING-BASED ML METHODS WITH AN AUTOMATED CLASSIFIER

In order to assess the difficulty of the image classification problem presented to participants and to evaluate the potential of hybrid human-ML approaches, we developed a deep learning image classification approach that leverages large training datasets. Our classifier is based on ResNet-50, a popular variant of ResNet architecture (He et al., 2015), which has shown very good performance on multiple image classification tasks. It has been extensively used by the computer vision research community and adopted as a baseline architecture in many studies done over the last few years (Bello et al., 2021).

For training the classifier, we generated a balanced dataset of 100 k samples, with 10 k samples set aside as the validation set and the rest used as the training set. The images are representative of an even mixture of the difficulty classes used to generate Experiment Sets C and D. We trained and evaluated the performance of the network using training set sizes ranging from 10 k samples to 90 k samples, increasing the training set size by 10 k every iteration, totaling nine different training sessions. Each training session was started from the previous session's best-performing checkpoint of the network and the corresponding optimization state and continued for 35 epochs. See **Supplementary Material** for a complete description of the ResNet classifier used as well as a detailed analysis of its performance.

We emphasize that this work does not aim to advance the state-of-the-art results for automated image classification. Instead, the focus of the automated classification method is to explore the benefits and limitations of a hybrid method introduced herein that integrates the outputs of a well-known deep neural network into the crowdsourcing-based classification methods. In particular, the proposed method uses the output of the automated classifier as an additional feature of the featured ML methods. **Table 6** summarizes the results for the small imbalanced test sets used in Experiment Sets C and D as the training set grows larger. Due to the imbalanced nature of these test sets, this table and the rest of the analysis focus on F_1-score, false-negative rate (FNR), and area under the ROC curve (AUC). Before proceeding, it is worthwhile to mention two additional points regarding the values presented in the table. First, the input elicitation RC represents the probability value of positive classification obtained from the automated classifier when used as a feature. For example, BCE-RC indicates that both the binary elicitation inputs and the probability scores from the ResNet-50 were used as features for the ML classifiers. Second, the Combined Set C&D is created by merging the data from Experiment Sets C and D, thereby effectively doubling the size of the training set relative to the individual experiment sets.

Table 6 marks in bold those cases in which the performance of the hybrid method according to a given metric is better than both the completely automated approach (ResNet-50) and the results achieved by the crowdsourcing-based ML methods (according to the best input combination). As expected, when the ResNet-50 performance is poor, using its output as a feature hurts the overall results. Conversely, when the ResNet-50 performance is near perfect, it is difficult to improve upon its performance by adding information obtained from the crowd. However, apart from those extremes, exploiting the output of the ResNet-50 is beneficial in most cases, particularly regarding F_1-score and AUC.

The proposed hybrid methods, which use the results from the automated classifier as an additional input feature for the crowdsourcing-based ML methods, exhibited a robust performance. They attained maximum F_1-scores of 0.98, 0.96 0.97 and minimum FNRs of 0.04, 0.08, 0.06 for Experiment Set C, D, and Combined Set C&D, respectively, all of which represent significant improvements over what crowdsourcing-based methods achieved on a standalone basis. While these top results were associated with the automated classifier training set of 90k samples, impressive results were obtained using smaller datasets for Combined Set C&D, compared to Experiment Set C and D separately. As an example, incorporating the output of the automated classifier trained on 50k samples with the crowdsourcing-based methods for Combined Set C&D improved the F_1-score significantly (see **Tables 5**, **6**). However, the hybrid approach did not show better results for Experiment Sets C and D separately over the same training set size in some cases. This can be explained by the fact that Experiment Sets C and D have fewer data points than Combined Set C&D. This attests that, while crowdsourcing-based methods supplemented with the outputs of the automated classifier perform very well on small datasets, too few data points can negatively affect the performance of the hybrid approach.

7. DISCUSSION

This section highlights key observations related to the research questions, along with the limitations of the study. The experiment results demonstrate that supplementing binary choice elicitation with other forms of inputs can generate better classifiers. When the training sets is small, incorporating binary labels along with confidence values regarding these responses within any of the four ML classifiers tested in this work generated more dependable results for datasets of varying levels of difficulty. These diverse inputs also helped improve other performance metrics such as AUC values, which measure an ML model's capability to distinguishing between labels. While voting methods had a rather poor performance with respect to FNRs, a simple parametric modification (i.e., changing the threshold value) was shown to significantly reduce these values with comparatively small increases to FPRs. When the training sets is larger, integrating the inputs from the automated classifier with the crowdsourcing-based ML methods decreased FNRs even further. Those methods achieved near-perfect FNRs thanks to a large dataset that was used to train the automated classifier. The F_1-score was also improved significantly through this hybrid approach. Although smaller training sets of 50k samples slightly reduced the performance of the automated classifier, the numbers were still better than those obtained by standalone crowdsourcing-based methods. Altogether, the results demonstrate that including diverse inputs as features within an ML classifier, it is possible to obtain better classifications at a relatively low cost.

The methodology for aggregating crowd information to improve image classification outcomes presented in this paper could have wide-ranging applications. Through suitable adaptations and enhancements, it could be applied for various types of real-world screening tasks, such as inspecting luggage at travel checkpoints (e.g., airports, metro), X-ray imaging for medical diagnosis, online image labeling, AI model training using CAPTCHAs, etc. Moreover, the image classification problem featured herein is a special case of the overall participant information aggregation problem; therefore, the findings in this paper could be extended to various other classification applications that utilize the wisdom of the crowd concept.

The presented studies admittedly have some limitations. For starters, the approach used to filter "insincere participants" was relatively simple. To obtain a better quality dataset, future studies will seek to deploy more sophisticated quality control techniques for filtering out unreliable or poor quality participants, e.g., using Honeypot questions (Mortensen et al., 2017). A second limitation is that the synthetic images generated for this work have certain characteristics that may overly benefit automated classification methods but may not generalize to various real-world situations. It is possible, for example, that the images might

TABLE 6 | Performance analysis of Crowdsourcing-based ML methods with expanded inputs from ResNet-50.

Input Elicitations	Size of dataset	ResNet50			KNN			LR			RF			SVM-Linear		
		F_1	FNR	AUC	F_1	FNR	AUC	F_1	FNR	AUC	F_1	FNR	AUC	F_1	FNR	AUC
Experiment Set C																
BCE-CE-SE-PDE*	–	–	–	–	0.81	0.29	0.92	0.81	0.29	0.86	0.81	0.29	0.90	0.81	0.29	0.86
RC		0.36	0.21	0.67	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	10k	–	–	–	0.73	0.38	0.85	**0.82**	0.25	**0.93**	0.70	0.38	0.89	0.75	0.38	0.92
BCE-CE-RC		–	–	–	0.70	0.42	0.89	0.77	0.25	0.88	0.74	0.33	0.86	0.80	0.33	0.91
RC		0.71	0.29	0.92	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	30k	–	–	–	0.77	0.38	0.83	0.75	**0.25**	0.92	0.78	0.33	0.89	0.76	0.33	0.92
BCE-CE-RC		–	–	–	0.75	0.38	0.81	0.73	**0.25**	0.91	0.78	0.33	0.88	0.80	0.33	**0.93**
RC		0.87	0.04	0.99	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	50k	–	–	–	0.80	0.25	0.95	**0.90**	0.04	0.97	0.84	0.21	0.97	**0.88**	0.13	0.98
BCE-CE-RC		–	–	–	0.82	0.25	0.92	**0.90**	0.04	0.98	0.84	0.21	0.97	**0.88**	0.13	0.97
RC		0.90	0.08	0.99	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	70k	–	–	–	**0.91**	0.13	0.98	**0.92**	**0.04**	0.99	**0.93**	0.13	0.98	**0.94**	**0.04**	0.99
BCE-CE-RC		–	–	–	**0.91**	0.13	0.98	0.88	**0.04**	**1.00**	**0.93**	0.13	0.98	**0.94**	**0.04**	0.99
RC		0.96	0.00	1.00	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	90k	–	–	–	**0.98**	0.04	0.98	0.94	0.04	0.96	**0.98**	0.04	0.97	**0.98**	0.04	0.99
BCE-CE-RC		–	–	–	**0.98**	0.04	0.98	0.9	0.04	0.97	**0.98**	0.04	0.97	**0.98**	0.04	0.99
Experiment Set D																
BCE-CE*	–	–	–	–	0.59	0.58	0.76	0.54	0.42	0.83	0.63	0.50	0.79	0.67	0.50	0.87
RC		0.17	0.33	0.62	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	10k	–	–	–	0.59	0.58	0.67	0.44	**0.25**	0.87	0.44	0.67	0.73	0.11	0.42	0.78
BCE-CE-RC		–	–	–	0.56	0.58	0.69	0.43	0.33	0.84	0.63	0.50	0.78	0.63	0.50	0.86
RC		0.50	0.42	0.87	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	30k	–	–	–	0.50	0.67	0.67	0.43	**0.33**	0.87	0.59	0.58	0.74	0.63	0.50	0.85
BCE-CE-RC		–	–	–	0.50	0.67	0.64	0.47	0.42	**0.88**	0.56	0.58	0.8	0.67	0.50	0.87
RC		0.79	0.08	0.96	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	50k	–	–	–	0.74	0.42	0.90	0.71	0.17	0.91	0.70	0.42	0.88	**0.80**	0.17	0.96
BCE-CE-RC		–	–	–	0.70	0.42	0.91	0.69	0.17	0.90	0.74	0.42	0.86	**0.80**	0.17	0.91
RC		0.83	0.17	0.98	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	70k	–	–	–	**0.91**	0.17	0.96	**0.88**	**0.08**	0.92	**0.91**	0.17	0.94	**0.96**	**0.08**	0.92
BCE-CE-RC		–	–	–	**0.91**	0.17	0.96	**0.88**	**0.08**	0.92	**0.91**	0.17	0.93	**0.96**	**0.08**	0.92
RC		0.96	0.08	0.98	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	90k	–	–	–	0.96	0.08	0.96	0.92	0.08	0.94	0.91	0.17	0.94	0.96	0.08	0.95
BCE-CE-RC		–	–	–	0.96	0.08	0.96	0.92	0.08	0.95	0.91	0.17	0.94	0.96	0.08	0.92
Combined Set C&D																
BCE-CE*	–	–	–	–	0.68	0.47	0.83	0.73	0.33	0.9	0.72	0.42	0.85	0.76	0.39	0.9
RC		0.27	0.25	0.65	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	10k	–	–	–	0.67	0.5	0.83	0.64	0.25	0.88	0.67	0.39	0.86	0.71	0.39	**0.91**
BCE-CE-RC		–	–	–	0.69	0.44	0.86	0.68	0.25	0.9	0.69	0.44	0.84	0.76	0.39	**0.91**
RC		0.63	0.33	0.90	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	30k	–	–	–	0.71	0.42	0.87	0.66	**0.25**	**0.92**	**0.79**	**0.31**	**0.94**	0.72	0.42	**0.91**
BCE-CE-RC		–	–	–	0.72	0.42	0.84	0.64	**0.28**	**0.92**	0.75	0.39	**0.91**	0.72	0.42	**0.93**
RC		0.84	0.06	0.98	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	50k	–	–	–	**0.87**	0.19	0.96	**0.86**	0.08	0.97	**0.91**	0.11	0.96	**0.85**	0.14	0.97
BCE-CE-RC		–	–	–	**0.86**	0.22	0.94	**0.86**	0.08	0.96	**0.86**	0.22	0.96	**0.85**	0.14	0.98
RC		0.88	0.11	0.99	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	70k	–	–	–	**0.96**	**0.08**	0.97	**0.92**	**0.06**	0.94	**0.96**	**0.08**	0.96	**0.93**	**0.06**	0.97
BCE-CE-RC		–	–	–	**0.93**	**0.08**	0.97	**0.89**	**0.06**	0.97	**0.96**	**0.08**	0.96	**0.92**	**0.06**	0.97
RC		0.96	0.03	0.99	–	–	–	–	–	–	–	–	–	–	–	–
BCE-RC	90k	–	–	–	**0.97**	0.06	0.97	0.93	0.06	0.98	**0.97**	0.06	0.96	**0.97**	0.06	0.98
BCE-CE-RC		–	–	–	**0.97**	0.06	0.97	0.92	0.06	0.94	**0.97**	0.06	0.97	**0.97**	0.06	0.98

**Denotes the input combinations that achieved the best performance among the Crowdsourcing-based ML methods. Bold values denote cases where hybrid method outperforms both the Resnet-50 classifier and the Crowdsourcing-based ML methods.*

have tiny consistent details that are not visible to human eyes due to the nature of the image generation process. In that case, the automated classification method had an unfair advantage of exploiting those details to improve performance effectively. Future studies will assess the featured methods on more realistic datasets drawn from other practical contexts.

8. CONCLUSION

Although crowdsourcing methods have been productive in image classification, they do not tap into the full potential of the wisdom of the crowd in one important respect. These methods have largely overlooked the fact that difficult tasks can be amplified to elicit and integrate multiple inputs from each participant; an easy-to-implement option, for example, is eliciting the level of confidence in one's binary response. This paper investigates how different types of information can be utilized with machine learning to enhance the capabilities of crowdsourcing-based classification. It makes four main contributions. First, it introduces a systematic synthetic image generation process that can be used to create image classification tasks of varying difficulty. Second, it demonstrates that while reported confidence in one's response does not significantly raise the performance of voting methods, this intuitive form of input can enhance the performance of machine learning methods, particularly when smaller training datasets are available. Third, it explains how aggregation methods can be adapted to prioritize other metrics of interest of image classification (e.g., reduced false-negative rates). Fourth, it demonstrates that under the right circumstances, automated classifiers can significantly improve classification performance when integrated with crowdsourcing-based methods.

The code used to generate the synthetic images can be found at https://github.com/O-ARE/2D-Image-Generation-HCOMP. In addition, the code used to train and evaluate the automated classifier can be found at https://github.com/O-ARE/2d-image-classification.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article can be made available by the authors upon request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Tiffany Dunning, IRB Coordinator, Arizona State University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RY, AE, OF, MH, and JG contributed to the conception and design of the study. HB organized the database. RY, JG, and MH performed the computational analysis. RY wrote the first draft of the manuscript. MH, JG, and HB wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This material was based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-05-00, which all authors gratefully acknowledge. The lead PI of the project (AE) and two of the students (JG and HB) also gratefully acknowledge support from the National Science Foundation under Award Number 1850355. An earlier, shorter version of this paper and a smaller subset of the results featured herein have been published in Yasmin et al. (2021) and presented in the 9th AAAI Conference on Human Computation and Crowdsourcing.

ACKNOWLEDGMENTS

The authors thank all participants in this study, which received institutional IRB approval prior to deployment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.848056/full#supplementary-material

REFERENCES

Assiri, Y. (2020). Stochastic optimization of plain convolutional neural networks with simple methods. *arXiv [Preprint] arXiv*:2001.08856. doi: 10.48550/arXiv.2001.08856

Barrington, L., Ghosh, S., Greene, M., Har-Noy, S., Berger, J., Gill, S., et al. (2012). Crowdsourcing earthquake damage assessment using remote sensing imagery. *Ann. Geophys*. 54. doi: 10.4401/ag-5324

Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., et al. (2021). Revisiting resnets: improved training and scaling strategies. *arXiv [Preprint] arXiv*:2103.07579. doi: 10.48550/arXiv.2103.07579

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (2016). *Handbook of Computational Social Choice*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781107446984

Chang, J. C., Amershi, S., and Kamar, E. (2017). "Revolt: collaborative crowdsourcing for labeling machine learning datasets," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY), 2334–2346. doi: 10.1145/3025453.3026044

Cheplygina, V., de Bruijne, M., and Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* 54, 280–296. doi: 10.1016/j.media.2019.03.009

Cheplygina, V., and Pluim, J. P. (2018). "Crowd disagreement about medical images is informative," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, eds D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, et al. (Granada: Springer), 105–111. doi: 10.1007/978-3-030-01364-6_12

Christoforou, E., Fernández Anta, A., and Sánchez, A. (2021). An experimental characterization of workers' behavior and accuracy in crowdsourced tasks. *PLoS ONE* 16:e0252604. doi: 10.1371/journal.pone.0252604

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *ArXiv, abs/2106.04803*. doi: 10.48550/arXiv.2106.04803

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL), 248–255. doi: 10.1109/CVPR.2009.5206848

Dodge, S., and Karam, L. (2017). "A study and comparison of human and deep learning recognition performance under visual distortions," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)* (Vancouver, BC), 1–7. doi: 10.1109/ICCCN.2017.8038465

Eickhoff, C. (2018). "Cognitive biases in crowdsourcing," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (New York, NY), 162–170. doi: 10.1145/3159652.3159654

Escobedo, A. R., Moreno-Centeno, E., and Yasmin, R. (2022). An axiomatic distance methodology for aggregating multimodal evaluations. *Inform. Sci*. 590, 322–345. doi: 10.1016/j.ins.2021.12.124

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD, Vol. 96* (Portland, OR), 226–231.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis*. 88, 303–338. doi: 10.1007/s11263-009-0275-4

Foody, G., See, L., Fritz, S., Moorthy, I., Perger, C., Schill, C., and Boyd, D. (2018). Increasing the accuracy of crowdsourced information on land cover *via* a voting procedure weighted by information inferred from the contributed data. *ISPRS Int. J. Geo-Inform*. 7:80. doi: 10.3390/ijgi7030080

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv [Preprint] arXiv*:1706.06969. doi: 10.48550/arXiv.1706.06969

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., et al. (2020). Expert-augmented machine learning. *Proc. Natl. Acad. Sci. U.S.A*. 117, 4571–4577. doi: 10.1073/pnas.1906831117

Görzen, T., Laux, F., et al. (2019). *Extracting the Wisdom From the Crowd: A Comparison of Approaches to Aggregating Collective Intelligence*. Technical Report, Paderborn University, Faculty of Business Administration and Economics.

Griffin, D., and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cogn. Psychol*. 24, 411–435. doi: 10.1016/0010-0285(92)90013-R

Grofman, B., Owen, G., and Feld, S. L. (1983). Thirteen theorems in search of the truth. *Theory Decis*. 15, 261–278. doi: 10.1007/BF00125672

Gurari, D., Theriault, D., Sameki, M., Isenberg, B., Pham, T. A., Purwada, A., et al. (2015). "How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms," in *2015 IEEE Winter Conference on Applications of Computer Vision* (Austin, TX), 1169–1176. doi: 10.1109/WACV.2015.160

Hamada, D., Nakayama, M., and Saiki, J. (2020). Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cogn. Res*. 5, 1–15. doi: 10.1186/s41235-020-00248-z

Hara, K., Le, V., and Froehlich, J. (2012). "A feasibility study of crowdsourcing and google street view to determine sidewalk accessibility," in *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY), 273–274. doi: 10.1145/2384916.2384989

Hastie, R., and Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychol. Rev*. 112:494. doi: 10.1037/0033-295X.112.2.494

He, J., van Ossenbruggen, J., and de Vries, A. P. (2013). "Do you need experts in the crowd? A case study in image annotation for marine biology," in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (Lisbon), 57–60.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv [Preprint] arXiv*:1512.03385. doi: 10.1109/CVPR.2016.90

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., et al. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *Eur. J. Cancer* 120, 114–121.

Hsing, P.-Y., Bradley, S., Kent, V. T., Hill, R. A., Smith, G. C., Whittingham, M. J., et al. (2018). Economical crowdsourcing for camera trap image classification. *Remote Sens. Ecol. Conserv*. 4, 361–374. doi: 10.1002/rse2.84

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). "Quality management on amazon mechanical Turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation* (Washington, DC), 64–67. doi: 10.1145/1837885.1837906

Irshad, H., Oh, E.-Y., Schmolze, D., Quintana, L. M., Collins, L., Tamimi, R. M., et al. (2017). Crowdsourcing scoring of immunohistochemistry images: evaluating performance of the crowd and an automated computational method. *Sci. Rep*. 7, 1–10. doi: 10.1038/srep43286

Jeannin, S., and Bober, M. (1999). *Description of Core Experiments for MPEG-7 Motion/Shape*. MPEG-7, ISO/IEC/JTC1/SC29/WG11/MPEG99 N, 2690.

Karger, D. R., Oh, S., and Shah, D. (2011). "Iterative learning for reliable crowdsourcing systems," in *Neural Information Processing Systems* (Granada).

Kemmer, R., Yoo, Y., Escobedo, A., and Maciejewski, R. (2020). "Enhancing collective estimates by aggregating cardinal and ordinal inputs," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 8* (New York, NY), 73–82.

Khattak, F. K., and Salleb-Aouissi, A. (2011). "Quality control of crowd labeling through expert evaluation," in *Proceedings of the NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds, Vol. 2* (Sierra Nevada), 5.

Koh, W. L., Kaliappan, J., Rice, M., Ma, K.-T., Tay, H. H., and Tan, W. P. (2017). "Preliminary investigation of augmented intelligence for remote assistance using a wearable display," in *TENCON 2017-2017 IEEE Region 10 Conference* (Penang), 2093–2098. doi: 10.1109/TENCON.2017.8228206

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychol. Rev*. 119:80. doi: 10.1037/a0025648

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems, Vol. 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Penang: Curran Associates, Inc.).

Li, Q., and Varshney, P. K. (2017). "Does confidence reporting from the crowd benefit crowdsourcing performance?" in *Proceedings of the 2nd International Workshop on Social Sensing* (New York, NY), 49–54. doi: 10.1145/3055601.3055607

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., and Hertwig, R. (2020). How the "wisdom of the inner crowd" can boost accuracy of confidence judgments. *Decision* 7:183. doi: 10.1037/dec0000119

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *J. Pers. Soc. Psychol*. 107:276. doi: 10.1037/a0036677

Mao, A., Procaccia, A. D., and Chen, Y. (2013). "Better human computation through principled voting," in *AAAI* (Bellevue, WA).

Matoulkova, B. K. (2017). *Wisdom of the crowd: comparison of the CWM, simple average and surprisingly popular answer method* (Master's thesis). Erasmus University Rotterdam, Rotterdam, Netherlands.

Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., et al. (2012). Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. *PLoS ONE* 7:e37245. doi: 10.1371/journal.pone.0037245

McDaniel, P., Papernot, N., and Celik, Z. B. (2016). Machine learning in adversarial settings. *IEEE Secur. Privacy* 14, 68–72. doi: 10.1109/MSP.2016.51

Meyen, S., Sigg, D. M., von Luxburg, U., and Franz, V. H. (2021). Group decisions based on confidence weighted majority voting. *Cogn. Res*. 6, 1–13. doi: 10.1186/s41235-021-00279-0

Mitry, D., Peto, T., Hayat, S., Morgan, J. E., Khaw, K.-T., and Foster, P. J. (2013). Crowdsourcing as a novel technique for retinal fundus photography classification: analysis of images in the epic norfolk cohort on behalf of the ukbiobank eye and vision consortium. *PLoS ONE* 8:e71154. doi: 10.1371/journal.pone.0071154

Mitry, D., Zutis, K., Dhillon, B., Peto, T., Hayat, S., Khaw, K.-T., et al. (2016). The accuracy and reliability of crowdsource annotations of digital retinal images. *Transl. Vis. Sci. Technol*. 5:6. doi: 10.1167/tvst.5.5.6

Mora, D., Zimmermann, R., Cirqueira, D., Bezbradica, M., Helfert, M., Auinger, A., and Werth, D. (2020). "Who wants to use an augmented reality shopping assistant application?" in *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications - WUDESHI-DR* (SciTePress), 309–318. doi: 10.5220/0010214503090318

Mortensen, M. L., Adam, G. P., Trikalinos, T. A., Kraska, T., and Wallace, B. C. (2017). An exploration of crowdsourcing citation screening for systematic reviews. *Res. Synthes. Methods* 8, 366–386. doi: 10.1002/jrsm.1252

Nguyen, T. B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., et al. (2012). Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. *Radiology* 262, 824–833. doi: 10.1148/radiol.11110938

Nowak, S., and Rüger, S. (2010). "How reliable are annotations *via* crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the International Conference on Multimedia Information Retrieval* (New York, NY), 557–566. doi: 10.1145/1743384.1743478

Oosterman, J., Nottamkandath, A., Dijkshoorn, C., Bozzon, A., Houben, G.-J., and Aroyo, L. (2014). "Crowdsourcing knowledge-intensive tasks in cultural heritage," in *Proceedings of the 2014 ACM Conference on Web Science* (New York, NY), 267–268. doi: 10.1145/2615569.2615644

Oyama, S., Baba, Y., Sakurai, Y., and Kashima, H. (2013). "Accurate integration of crowdsourced labels using workers' self-reported confidence scores," in *Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing).

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016). "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (Saarbrücken), 372–387. doi: 10.1109/EuroSP.2016.36

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature* 541:532. doi: 10.1038/nature21054

Ralph, R. Mpeg-7 core experiment ce-shape-1 test set. (1999). Available online at: https://dabi.temple.edu/external/shape/MPEG7/dataset.html

Rankin, W. L., and Grube, J. W. (1980). A comparison of ranking and rating procedures for value system measurement. *Eur. J. Soc. Psychol.* 10, 233–246. doi: 10.1002/ejsp.2420100303

Rasp, S., Schulz, H., Bony, S., and Stevens, B. (2020). Combining crowdsourcing and deep learning to explore the mesoscale organization of shallow convection. *Bull. Am. Meteorol. Soc.* 101, E1980–E1995. doi: 10.1175/BAMS-D-19-0324.1

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Saab, F., Elhajj, I. H., Kayssi, A., and Chehab, A. (2019). Modelling cognitive bias in crowdsourcing systems. *Cogn. Syst. Res.* 58, 1–18. doi: 10.1016/j.cogsys.2019.04.004

Saha Roy, T., Mazumder, S., and Das, K. (2021). Wisdom of crowds benefits perceptual decision making across difficulty levels. *Sci. Rep.* 11, 1–13. doi: 10.1038/s41598-020-80500-0

Salek, M., Bachrach, Y., and Key, P. (2013). "Hotspotting-a probabilistic graphical model for image object localization through crowdsourcing," in *Twenty-Seventh AAAI Conference on Artificial Intelligence* (Bellevue, WA).

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian Joint Conference on Artificial Intelligence* (Hobart, TAS: Springer), 1015–1021. doi: 10.1007/11941439_114

Stevens, B., Bony, S., Brogniez, H., Hentgen, L., Hohenegger, C., Kiemle, C., et al. (2020). Sugar, gravel, fish and flowers: mesoscale cloud patterns in the trade winds. *Q. J. R. Meteorol. Soc.* 146, 141–152. doi: 10.1002/qj.3662

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., and Packer, C. (2015). Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Sci. Data* 2, 1–14. doi: 10.1038/sdata.2015.26

Tan, M., and Le, Q. (2019). "EfficientNet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning* (Long Beach, CA), 6105–6114.

Xu, A., and Bailey, B. (2012). "A reference-based scoring model for increasing the findability of promising ideas in innovation pipelines," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (New York, NY), 1183–1186. doi: 10.1145/2145204.2145380

Yasmin, R., Grassel, J. T., Hassan, M. M., Fuentes, O., and Escobedo, A. R. (2021). "Enhancing image classification capabilities of crowdsourcing-based methods through expanded input elicitation," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 9* (Virtual), 166–178.

Yi, S. K. M., Steyvers, M., Lee, M. D., and Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cogn. Sci.* 36, 452–470. doi: 10.1111/j.1551-6709.2011.01223.x

Yoo, Y., Escobedo, A., and Skolfield, K. (2020). A new correlation coefficient for comparing and aggregating non-strict and incomplete rankings. *Eur. J. Oper. Res.* 285, 1025–1041. doi: 10.1016/j.ejor.2020.02.027

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2021). Scaling vision transformers. *arXiv [Preprint] arXiv*:2106.04560. doi: 10.48550/arXiv.2106.04560

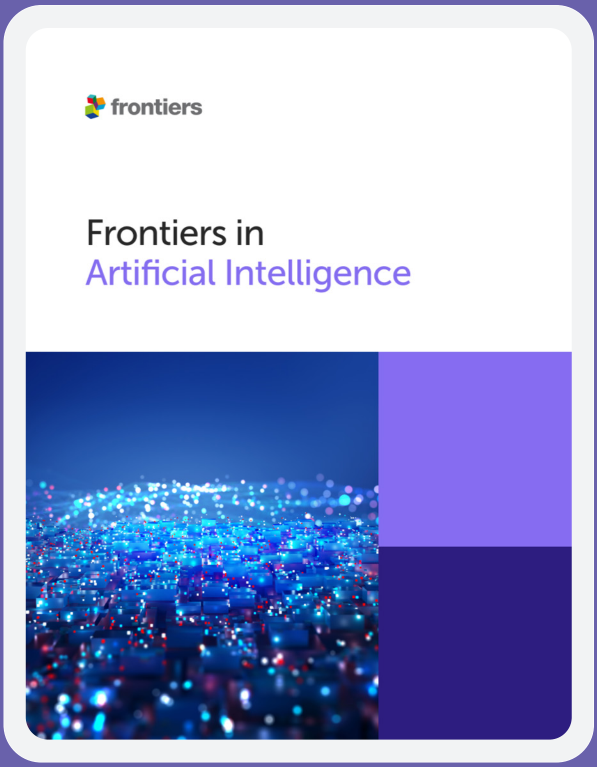
Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. *Adv. Neural Inform. Process. Syst.* 27, 487–495. doi: 10.1101/265918

Zhou, N., Siegel, Z. D., Zarecor, S., Lee, N., Campbell, D. A., Andorf, C. M., et al. (2018). Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. *PLoS Comput. Biol.* 14:e1006337. doi: 10.1371/journal.pcbi.1006337

Author Disclaimer: The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or of the National Science Foundation.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



frontiers

Frontiers in
Artificial Intelligence