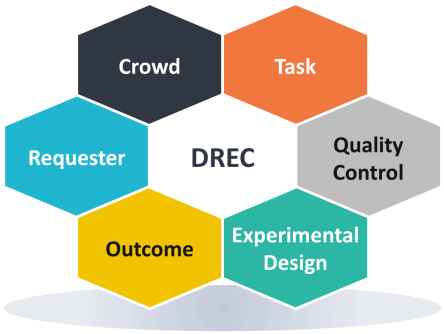

Figure 1: The datasheet for reporting experiments in crowdsourcing involves six high-level concepts: the crowdsourcing workers, the actual crowdsourced task, quality control mechanisms, the design and outcome of the experiment, and the task requester.

DREC: towards a Datasheet for Reporting Experiments in Crowdsourcing

Jorge Ramírez
jorge.ramirezmedina@unitn.it
University of Trento
Trento, Italy

Marcos Baez
Université Claude Bernard Lyon 1
Lyon, France

Fabio Casati
Tomsk Polytechnic University
Tomsk, Russia

Luca Cernuzzi
Catholic University of Asunción
Asunción, Paraguay

Boualem Benatallah
University of New South Wales
Sydney, Australia

ABSTRACT

Factors such as instructions, payment schemes, platform demographics, along with strategies for mapping studies into crowdsourcing environments, play an important role in the reproducibility of results. However, inferring these details from scientific articles is often a challenging endeavor, calling for the development of proper reporting guidelines. This paper makes the first steps towards this goal, by describing an initial taxonomy of relevant attributes for crowdsourcing experiments, and providing a glimpse into the state of reporting by analyzing a sample of CSCW papers.

CSCW '20 Companion, October 17–21, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8059-1/20/10.
https://doi.org/10.1145/3406865.3418318

Table 1: Proposed taxonomy of crowdsourcing experiments attributes.

Dimension	Description
Crowd	Crowdsourcing workers that participate in the experiment. **Attributes**: *reputation*, *environment*, *population sampling*.
Task(s)	The actual crowdsourcing task(s) shown to workers. **Attributes**: *type*, *modularity*, *task interface*, *instructions*, *reward strategy*, *time alloted*, *task assignment*.
Quality control	The mechanisms to guard the quality of the results. **Attributes**: *rejection criteria*, *votes per item*, *pre-task checks*, *training*, *in-task checks*, *post-task checks*, *dropouts prevention mechanisms*, *dropout rate*.
Experimental design	The design of the experiment(s) involving crowdsourcing workers. **Attributes**: *research questions*, *input dataset*, *experimental variables*, *random assignments*, *experimental conditions*, *synchronous*, *study design*, *execution dates*,*pilots*, *returning workers*.
Outcome	The results of the experiment. **Attributes**: *data analysis*, *discarded data*, *excluded participants*, *number of participants*, *demographics*, *number of contributions*, *output dataset*.
Requester	The individual or institution that runs the study. **Attributes**: *platforms used*, *implemented features*, *fair compensation*, *requester-worker interactions*, *privacy*, *informed consent*, *ethical approvals*.

[1]Supplementary material can be found at https://tinyurl.com/DREC-cscw-poster

BACKGROUND & MOTIVATION

Currently, there are no standardized guidelines for properly reporting crowdsourcing experiments. This lack of reporting procedure can be potentially problematic due to the many details that constitute crowdsourcing projects, which can potentially damage the validity and reproducibility of results [19].

Poor practices in reporting is a serious concern in science, and crowdsourcing is no exception [4]. Researchers have shown how elements such as instructions [13, 28], interface [20, 21, 24], payments [11, 27], allotted time [14, 15], or the underlying platform demographics [19, 27] can potentially affect the outcomes of crowdsourcing projects. This potential variability of the results and the fact that researchers need to overcome the limitations of crowdsourcing platforms to run controlled experiments [8, 12, 16] emphasizes the need for a standardized guideline for reporting crowdsourcing.

There is abundant literature that offers guidelines for leveraging crowdsourcing in general [3, 10, 23] and for experiments in particular [5, 16, 22]. However, how to report crowdsourcing experiments have received little attention [18]. Inspired by recent work from the machine learning community [9, 17], we present our ongoing work aiming at a datasheet for reporting crowdsourcing experiments.

METHODOLOGY & PROPOSED TAXONOMY

We set to understand the main ingredients of crowdsourcing experiments and propose a taxonomy that captures this knowledge. With this taxonomy, we aim to derive a standardized procedure for reporting crowdsourcing experiments in the fields of computer science. We limited the focus to papers reporting user studies evaluating various aspects of task design and those focused on behavior understanding. This initial scope is motivated by the lack of support from crowdsourcing platforms when performing controlled experiments [8, 12, 16]. These limitations naturally translate into additional steps and details that researchers must communicate properly to aid reproducible research. Therefore, we considered this scope to produce an initial taxonomy that could apply to other crowdsourcing usages.

We got inspired by methodological approaches focused on standardizing the reporting of literature reviews [1, 26]. We build the taxonomy by relying on four sources: 1) guidelines for experimental research, 2) guidelines for crowdsourcing experiments, 3) task design features in crowdsourcing platforms (i.e., to support the deployment of experiments and outcomes reporting), and 4) papers reporting crowdsourcing experiments. We facilitate the related materials in the supplementary page [1].

We started with a set of papers as seed and used Google Scholar and Scopus to identify papers for points (1) and (2) in the list of sources, and considered the Toloka crowdsourcing platform for point (3). Individual attributes, as well as potential categorizations of these, were inferred and extracted from the guidelines of reference in a spreadsheet. The attributes were then analyzed and organized by two researchers to form an initial taxonomy consisting of six high-level concepts as depicted by Figure 1.

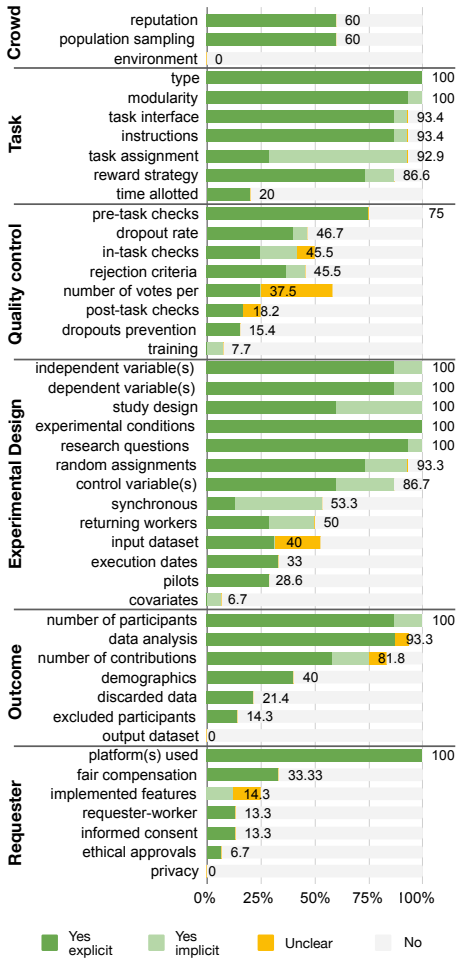

Figure 2: The state of reporting per attribute in the proposed taxonomy for the sample of CSCW papers from 2013 to 2019. Overall, the completeness level of the papers was between 31% and 67%, with 60% of the studies reporting at least half of the attributes in the taxonomy.

To collect papers reporting crowdsourcing experiments, point (4), we considered the Scopus database and iteratively refined a query to retrieve potentially relevant documents from major conferences. We narrowed down the list to those published in CSCW from 2013 to 2019, resulting in 72 papers. Two researchers screened the documents and identified 15 relevant papers.

The initial taxonomy was updated in an iterative process, as we examined the relevant papers reporting on crowdsourcing experiments. The information extraction was performed by two researchers in two phases, to test the taxonomy and identify possible attributes missing from the initial taxonomy. This process was performed starting from a random sample of five papers, and then processing the rest based on publication date until we reached saturation [25]. The researchers then met to discuss the final list, where organization, potential duplicates, and relevance were discussed and reached by consensus. Table 1 summarizes the taxonomy we propose in this paper, and a detailed version can be found in the supplementary page. It is the starting point for developing the datasheet.

A GLIMPSE ON THE STATE OF REPORTING

Two researchers annotated the selected sample of 15 papers by answering two basic questions for each attribute in the taxonomy: 1) *Can the attribute be derived from the paper?* (Completeness), and 2) *How is the attribute reported?* (Reporting). Two papers, different from the five random papers initially selected, were independently analyzed by two researchers to test and refine the level of agreement (93%). The remaining papers were distributed among the two researchers and annotated independently, with special cases being discussed and resolved by consensus.

By focusing on these two aspects, we aim to distill what attributes researchers currently communicate and how they report them (e.g., explicitly, implicitly). Notice that at this level of assessment, and stage in our project, we are setting the same weight to each attribute in our taxonomy. Naturally, for some experimental settings, some of the entries in our taxonomy may not be necessary or applicable. Therefore we are not judging the quality of the reporting. Figure 2 summarizes the results.

The variability of the underlying platform demographics [19, 27], and the environment in which workers perform the tasks [7] could affect the outcomes of crowdsourcing experiments. For the *Crowd* dimension, it can be noticed that 60% of the papers reported the mechanism used to assess the experience of workers (the *reputation*, for example, acceptance rate) and the criteria used to determine which workers participate in the study (the *population sampling*, e.g., demographics). In contrast, no paper reported the *environment* used by workers (e.g., the operating system or browser).

Task design is also known to impact crowdsourcing outcomes [21, 27, 28]. Around 93% of the papers reported the actual task interface and instructions. Screenshots showing the task interface were the most common style of reporting (9/14) and also references to papers or related systems (4/14). For the instructions, authors resorted to text descriptions (7/14) describing or providing excerpts and screenshots (6/14) in most cases alongside the task interface. Attributes such as task assignment (93%)

Table 2: Summary of insights.

- While task design aspects are largely reported, it is not clear whether screenshots can provide the necessary attributes to infer proper task designs, which ultimately can hurt reproducibility.
- Some attributes related to task design and quality control rely on features from the crowdsourcing platforms and go unreported. These implicit attributes might pose some problems when porting the task to a different platform.
- General experimental design aspects are well reported, but less so those concerning to crowdsourcing. We believe this is due to more established guidelines for the former in comparison to the latter.
- Information regarding the population and the resulting crowdsourcing datasets should be made available to foster reproducible research. Inherent human biases and those present in the underlying crowdsourcing platform plus the underreporting of these aspects could harm the replicability of a study.
- Ethical and privacy aspects need to be better reported. The ethics and fairness of crowd work have received much attention recently compared to the early days of crowdsourcing [2]. In this context, we found that some of the papers report that they set up the payment to meet a minimum wage, and only a small fraction of these reported an informed consent and ethical approval.

and reward strategy (87%) were also mostly derived from study design descriptions, although for task assignment implicitly from the selected platform. Very few papers reported time constraints (20%).

Quality control mechanisms are crucial to ensuring high-quality contributions [6]. The papers analyzed do not consistently report on the implemented mechanisms, the most acknowledged one being pre-task checks (e.g., pre-filtering workers based on task acceptance rate) in 75% of the papers, while the rest in less than half of the studies. Attributes in this dimension were reported as text. We notice that we only included instances where the quality controls were applicable (e.g., post-checks were not considered applicable in studies analyzing the quality of outcomes).

Experimental design aspects set the tone for the overall study setup, and are thus fundamental to a crowdsourcing experiment. Key general aspects are highly reported (RQs, study design, experimental variables), which were derived directly from text, tables and figures in the methods and results sections. In contrast, aspects more closely to crowdsourcing were less reported. At best, 50% of the papers reported input datasets, handling returning workers, execution dates, and whether pilots were run.

The *outcome* is important to understanding the results, verifying and replicating the crowdsourcing experiment. In this regard, participants' information was addressed with varying degrees, with all studies reporting on total numbers, but few on demographics (40%) or whether participants were discarded (14%). Most studies reported the number of collected contributions (82%) but not so the information on discarded data (21%). Surprisingly, papers did not share the resulting output dataset.

The details related to how the *requester* prepared and setup the experiment are equally important. Platforms were reported in all studies, but extensions to deal with the limited platform support for controlled experiments, reported in only 14% of the studies. The human aspect of the experiment preparation, dealing with fairness (33%), ethics (7%) and privacy (0%), went largely unreported.

DISCUSSION & ONGOING WORK

We highlighted the need for guidelines for reporting on crowdsourcing experiments and proposed an initial taxonomy of relevant attributes. With this taxonomy, we characterized the level of reporting in the CSCW community and derived some insights summarized in Table 2. Interestingly, high-level aspects of experimental design, and basic aspects of task design, were generally well covered by the analyzed studies. Other elements related to how experimental designs are mapped to crowdsourcing platforms were less addressed and represent potential threats to the validity of the results. As part of our ongoing work we are i) designing empirical studies to understand the effect of unreported factors on experiment outcomes, ii) further developing the initial taxonomy into a reporting tool, and iii) extending the scope of the reporting to other crowdsourcing tasks such as data collection.

ACKNOWLEDGMENTS

This work was supported by the Russian Science Foundation (Project No. 19-18-00282).

REFERENCES

[1] Kitchenham Barbara and Stuart Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering. 2 (01 2007).

[2] Natã Miccael Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. 543. https://doi.org/10.1145/3290605.3300773

[3] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980 arXiv:https://doi.org/10.1177/1745691610393980 PMID: 26162106.

[4] Michael Buhrmester, Sanaz Talaifar, and Samuel Gosling. 2018. An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. *Perspectives on Psychological Science* 13 (03 2018), 149–154. https://doi.org/10.1177/1745691617706516

[5] Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2013. Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers. *Behavior research methods* 46 (07 2013). https://doi.org/10.3758/s13428-013-0365-7

[6] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1 (2018), 7:1–7:40. https://doi.org/10.1145/3148148

[7] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017).

[8] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel W. Archambault, and Brian Fisher. 2015. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments - Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22-27, 2015, Revised Contributions*. 6–26. https://doi.org/10.1007/978-3-319-66435-4_2

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). arXiv:1803.09010 http://arxiv.org/abs/1803.09010

[10] Joseph Goodman and G. Paolacci. 2017. Crowdsourcing consumer research. *Journal of Consumer Research* 44 (06 2017), 196–210. https://doi.org/10.1093/jcr/ucx047

[11] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*. 419–429. https://doi.org/10.1145/2736277.2741102

[12] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April 5-10, 2008*. 453–456. https://doi.org/10.1145/1357054.1357127

[13] Aniket Kittur, Jeffrey V. Nickerson, Michael S. Bernstein, Elizabeth Gerber, Aaron D. Shaw, John Zimmerman, Matt Lease, and John J. Horton. 2013. The future of crowd work. In *Computer Supported Cooperative Work, CSCW 2013, San Antonio, TX, USA, February 23-27, 2013*. 1301–1318. https://doi.org/10.1145/2441776.2441923

[14] Ranjay A Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. 2016. Embracing error to enable rapid crowdsourcing. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 3167–3179.

[15] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[16] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (01 Mar 2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6

[17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. 220–229. https://doi.org/10.1145/3287560.3287596

[18] Nathaniel D. Porter, Ashton M. Verdery, and S. Michael Gaddis. 2020. Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PLOS ONE* 15, 6 (06 2020), 1–21. https://doi.org/10.1371/journal.pone.0233154

[19] Rehab Kamal Qarout, Alessandro Checco, Gianluca Demartini, and Kalina Bontcheva. 2019. Platform-Related Factors in Repeatability and Reproducibility of Crowdsourcing Tasks. In *HCOMP 2019*.

[20] Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. 2019. Crowdsourced dataset to study the generation and impact of text highlighting in classification tasks. *BMC Research Notes* 12, 1 (2019), 820. https://doi.org/10.1186/s13104-019-4858-z

[21] Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. 2019. Understanding the Impact of Text Highlighting in Crowdsourcing Tasks. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019*, Vol. 7. AAAI, 144–152.

[22] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011.* http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778

[23] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.* 859–866. http://www.lrec-conf.org/proceedings/lrec2014/summaries/497.html

[24] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 - May 01, 2014.* 3665–3674. https://doi.org/10.1145/2556288.2557155

[25] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in Qualitative Research: Exploring its Conceptualization and Operationalization. *Quality & quantity* 52, 4 (2018), 1893–1907.

[26] Larissa Shamseer, David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 349 (2015). https://doi.org/10.1136/bmj.g7647 arXiv:https://www.bmj.com/content/349/bmj.g7647.full.pdf

[27] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the 2011 ACM Conference on Computer Supported Cooperative Work, CSCW 2011, Hangzhou, China, March 19-23, 2011.* 275–284. https://doi.org/10.1145/1958824.1958865

[28] Meng-Han Wu and Alexander J. Quinn. 2017. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. In *HCOMP 2017*. https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15943