CrowdHub: Extending crowdsourcing platforms for the controlled evaluation of tasks designs

Jorge Ramírez,[1] **Simone Degiacomi,** [1] **Davide Zanella,** [1]
Marcos Baez, [1] **Fabio Casati,** [1] **Boualem Benatallah** [2]

[1]University of Trento, [2]UNSW Sydney

Abstract

We present CrowdHub, a tool for running systematic evaluations of task designs on top of crowdsourcing platforms. The goal is to support the evaluation process, avoiding potential experimental biases that, according to our empirical studies, can amount to 38% loss in the utility of the collected dataset in uncontrolled settings. Using CrowdHub, researchers can map their experimental design and automate the complex process of managing task execution over time while controlling for returning workers and crowd demographics, thus reducing bias, increasing utility of collected data, and making more efficient use of a limited pool of subjects.

Background & Motivation

A crucial aspect in running a successful crowdsourcing project is identifying an appropriate task design (Jain et al. 2017). Studies have shown that worker behavior can be influenced by different factors such as task design and presentation (Sampath, Rajeshuni, and Indurkhya 2014), allocated time to completion (Maddalena et al. 2016), pricing, and reward schemes (Difallah et al. 2014), human aspects in collaboration and individual biases (Drapeau et al. 2016; Eickhoff 2018), and even characteristics of the crowd marketplace and work environment (Gadiraju et al. 2017).

However, the potential size of the design space, along with the individual and environmental biases, and the limitations of crowdsourcing platforms, makes it difficult to systematically study tasks designs. As a result, workers still deal with poorly designed tasks (Gadiraju, Yang, and Bozzon 2017) that can affect their performance and introduce systematic biases (Faltings et al. 2014), producing undesirable results for requesters or deterring aggregation models from deriving the right answers (Kamar, Kapoor, and Horvitz 2015).

In this demo, we explore challenges that arise when evaluating multiple task designs, and we introduce CrowdHub a tool for running crowdsourcing experiments, offering features to overcome these issues.

Challenges in Evaluating Task Designs

We explored the challenges in evaluating task designs while studying the impact of highlighting support in text classification (Ramírez et al. 2019). The goal was to understand *if*, and *under what conditions*, highlighting text excerpts relevant to a given relevance question would improve worker performance. This required testing different highlighting conditions (of varying quality) against a baseline without highlighting, given different document sizes and datasets of different characteristics. The resulting experimental design featured a combination of *dataset* (3) x *document size* (3) x *highlighting conditions* (6) - a total of 54 configurations.

Crowdsourcing platforms such as Figure Eight (F8) offer the building blocks to design and run crowdsourcing tasks. In F8, this implies defining i) data units to classify, ii) gold data to use for quality control, iii) task design, including instructions, data to collect, assignment of units to workers, iv) the target population (country, channels, trust), and v) the cost per worker contribution. These features are suitable for running individual tasks, but less so when experimenting with different task designs with a limited pool of workers.

In order to identify and quantify the experimental bias in running an *uncontrolled* evaluation of task designs, we created individual tasks in F8 for a subset (1 dataset) of the experimental conditions, and ran them one after another, collecting a total of 6993 votes from 631 workers (16 tasks). The analysis of the results points to the following issues:

– **Recurrent workers**. While returning workers are desirable in any crowdsourcing task, they represent a potential source of bias in the context of task design evaluation, i.e., they might perform better, due to the *learning* effect. In our experiments, we observed a 38% of returning workers, who featured a lower completion time but not higher accuracy[1] (Figure 1A).

– **Condition crossover**. Returning workers can also land in a different experimental condition, which could modify their behavior and performance. From the 30% workers who crossed conditions (Figure 1A), we observed that switching between highlighting support and not support resulted in lower decision time, but that having had a bad highlighting support before can increase the time when dealing with good support - possibly due to the lack of trust in the support. Workers switching from support to no

---

[1]We noticed, however, that accuracy remained mostly unaffected by conditions and other factors across all our experiments, and it might have been less susceptible to the learning effect.
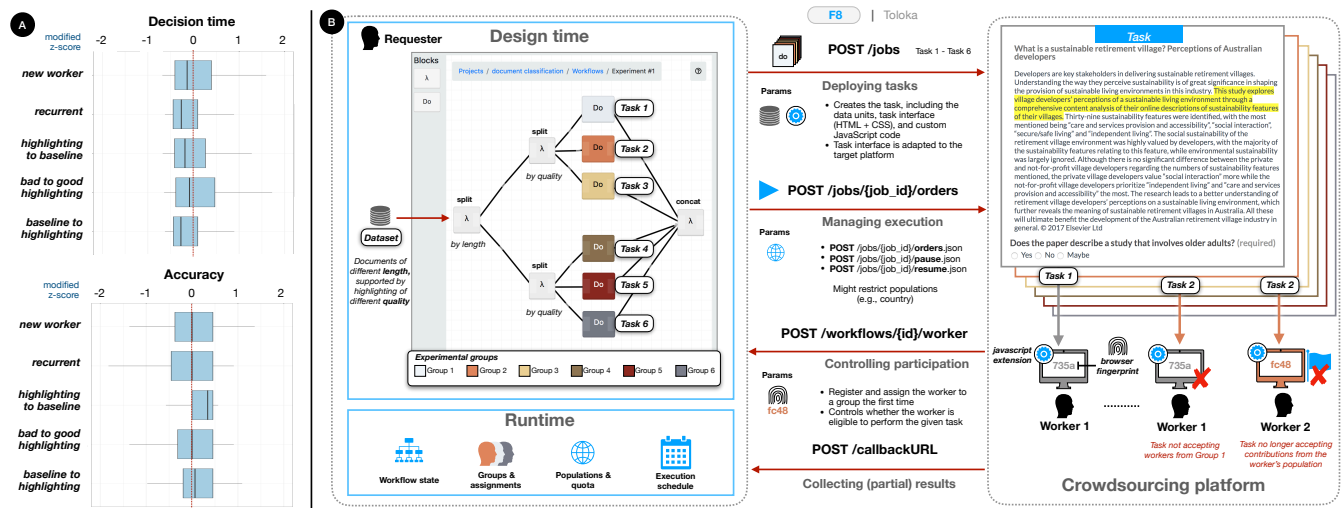
Figure 1: A) Decision time and accuracy for recurrent workers in the highlighting support experiment. Values are normalized to the distribution of *new workers* in each condition. B) Example workflow for a between-subjects design using CrowdHub.

support also featured higher accuracy than the new workers and those returning to the same condition.

– **Timezones**. Running the conditions at different times can introduce confounding factors that could hurt the comparison. For instance, we observed the worker performance in independent runs varied by different factors even between runs of the same condition (e.g., from 24s to 14s in decision time between a first and a second run considering only new workers) Thus collecting reliable and comparable results in this setting would require multiple runs over a long period of time.

– **Population demographics**. The pool of active workers is determined by the demographics of the crowdsourcing platform and the time an experiment runs. In running uncontrolled tasks, we observed a participation dominated by certain countries, which prevented more diverse population characteristics. For example, the top contributing countries provided 48.1% of the total judgements (Venezuela: 28.5%, Egypt: 11.8%, Ukraine: 7.8%).

Running a systematic comparison of task designs using the native building blocks of a crowdsourcing platform is thus a complex activity, susceptible to different types of experimental biases, which are costly to clean up (e.g., discarding 38% of the dataset) - a challenge many task designers and researchers face.

CrowdHub Platform

The above challenges motivated us to design and build a tool that extends the capabilities of crowdsourcing platforms for the purpose of task evaluation. CrowdHub is a system that sits on top of major crowd platforms, such as F8 and Toloka, and offers the building blocks to design and run controlled experiments using crowdsourcing.

We based our design on the following main ideas: workflows, eligibility control, population management and time sampling. *Workflows* allow requesters to set the foundation

for their experimental designs by defining the tasks and sequence of execution (sequential or parallel) — reducing issues in working with different active crowds. The *eligibility control* of tasks allows requesters to define the policy regarding returning workers and condition crossovers associated with the experimental design (between- or within-groups design). Through *population management* requesters can control for subgroups of workers dominating a dataset by assigning a specific quota, and *time sampling* helps in controlling for confounding factors by scheduling task execution over a period of time. Altogether, these features allow requester to be in control of their experimental design.

CrowdHub enables the entire task evaluation process, as shown in Figure 1. At *design time*, requesters use the workflow editor to define the experimental design, which includes the tasks (*Do* boxes) and the data flow (indicated by the arrows and the *lambda* functions describing data aggregation and partitioning). Experimental groups can also be defined and associated to one or more tasks, denoted in the diagram using different colors. When deploying the experiment, CrowdHub parses the workflow definition and creates the individual tasks in the target crowdsourcing platform with the associated data units and task design. At *run time*, the requestor can specify the population management strategy, and time sampling, if any, and the platform will launch, pause and resume the tasks, and constraint the workers access to tasks, accordingly. From a technical perspective, CrowdHub manages the interactions with the crowdsourcing platform through their public APIs and JavaScript extensions incorporated to the tasks, for additional metrics, worker control, and worker identification (browser fingerprinting) (Gadiraju and Kawase 2017).

In this demo we will show the entire evaluation lifecycle with the current version of CrowdHub [2], including workflow design, eligibility control and deployment on F8 and Toloka.

References

Difallah, D. E.; Catasta, M.; Demartini, G.; and Cudré-Mauroux, P. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

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*.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 162–170. ACM.

Faltings, B.; Jurca, R.; Pu, P.; and Tran, B. D. 2014. Incentives to counter bias in human computation. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*.

Gadiraju, U., and Kawase, R. 2017. Improving reliability of crowdsourced results by detecting crowd workers with multiple identities. In *International Conference on Web Engineering*, 190–205. Springer.

Gadiraju, U.; Checco, A.; Gupta, N.; and Demartini, G. 2017. Modus operandi of crowd workers: The invisible role of microtask work environments. *IMWUT* 1(3):49:1–49:29.

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, HT 2017, Prague, Czech Republic, July 4-7, 2017*, 5–14.

Jain, A.; Sarma, A. D.; Parameswaran, A. G.; and Widom, J. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. *PVLDB* 10(7):829–840.

Kamar, E.; Kapoor, A.; and Horvitz, E. 2015. Identifying and accounting for task-dependent bias in crowdsourcing. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA.*, 92–101.

Maddalena, E.; Basaldella, M.; De Nart, D.; Degl'Innocenti, D.; Mizzaro, S.; and Demartini, G. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Ramírez, J.; Baez, M.; Casati, F.; and Benatallah, B. 2019. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019*.

Sampath, H. A.; Rajeshuni, R.; and Indurkhya, B. 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.