Erschienen in: Evaluation in the crowd : crowdsourcing and human-centered experiments : Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22-27, 2015, revised contributions / Archambault, Daniel; Purchase, Helene C.; Hoßfeld, Tobias (Hrsg.). - Cham : Springer, 2017. - (Lecture Notes in Computer Science ; 10264). - S. 6-26. - ISBN 978-3-319-66434-7 https://dx.doi.org/10.1007/978-3-319-66435-4_2

Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd

Ujwal Gadiraju[1(✉)], Sebastian Möller[2], Martin Nöllenburg[3], Dietmar Saupe[4], Sebastian Egger-Lampl[5], Daniel Archambault[6], and Brian Fisher[7]

[1] Leibniz Universität Hannover, Hannover, Germany
gadiraju@L3S.de
[2] TU Berlin, Berlin, Germany
[3] Algorithms and Complexity Group, TU Wien, Vienna, Austria
[4] University of Konstanz, Konstanz, Germany
[5] Austrian Institute of Technology, Vienna, Austria
[6] Swansea University, Swansea, UK
[7] Simon Fraser University, Burnaby, Canada

1 Introduction

The notion of '*crowdsourcing*' was born nearly a decade ago in 2006[1], and since then the crowdsourcing paradigm has been widely adopted across a multitude of domains. Crowdsourcing solutions have been proposed and implemented to overcome obstacles that require human intelligence at a large scale. In the last decade there have been numerous applications of crowdsourcing both in research and practice (for example, [25,34]). In the realm of research, crowdsourcing has presented novel opportunities for qualitative and quantitative studies by providing a means to scale-up previously constrained laboratory studies and controlled experiments [44]. By exploiting crowdsourcing we can build ground truths for evaluation, access desired participants around the clock with a wide variety of demographics at will [31], and all within a short amount of time. This also comes with a number of challenges related to lack of control on research subjects and to data quality.

In this chapter, we first explore a few limitations of conducting experiments in the laboratory and those using crowdsourcing. We then deliberate on the typical requirements for human-centered experiments and the considerations necessary when transitioning from constrained laboratory experiments to the use of crowdsourcing. Previous works have established that crowdsourcing is a suitable means to acquire participants for social and behavioral science experiments [7,26,37,41] and have validated them for use in human-computer interaction and visualization experiments [24]. Several other domains are successfully

The original version of this chapter was revised. The affiliation of the third author was corrected. The erratum to this chapter is available at https://doi.org/10.1007/978-3-319-66435-4_8

[1] http://www.wired.com/2006/06/crowds/ last accessed 14 Jun 2017.

using crowdsourcing: Quality of Experience (QoE) assessment (see Chap. 7), software testing and software development, and network measurements. In this work, we identify the key factors of an experiment that determine its suitability to benefit from crowdsourcing. By juxtaposing the strengths and weaknesses of controlled laboratory experiments and those using crowdsourcing, determined through the inherent characteristics of the two paradigms, we present the reader with an overall understanding of the kinds of experiments that can benefit from the virtues of crowdsourcing and the cases that are less suitable for the same.

1.1 Limitations of Laboratory Experiments

Before crowdsourcing gained popularity as an alternative means for experimentation, human-centered experiments were traditionally conducted in a controlled laboratory setting. Despite a wealth of experimental findings resulting from such experiments, researchers also face several limitations and difficulties when preparing, running, and analyzing laboratory experiments. Many of the limitations are linked to the possible scale of the experiments. Often the pool of participants is constrained to a rather small and not necessarily representative group of subjects that are easily accessible to an experimenter, e.g., college students enrolled in the same program and required to participate in a number of experiments during their studies. This makes it difficult to generalize the experimental findings to larger and culturally or educationally more heterogeneous groups of the population. Scaling laboratory experiments to larger numbers and more representative groups of participants immediately results in a strong increase in cost for personnel and participant remuneration, as well as in the actual time required to prepare and run the experiment. Both factors may often be prohibitive, especially in an academic setting with limited funds and resources. Moreover, the artificially controlled environment in the laboratory, while advantageous, e.g., for excluding external confounding factors or testing specialized equipment, also leads to a limited ecological validity, as the experimental tasks might be performed differently by the participants in a real-life setting.

1.2 Limitations of Crowdsourcing Experiments

Although crowdsourcing evidently empowers us with an ability to run experiments using a large number of participants at a previously unmatched scale, there are a few concomitant pitfalls. Due to varying motivations of participants in the crowd (in both reward-based and to a lesser extent in altruistic crowdsourcing), quality control is a major challenge. Several prior works have addressed this issue [11, 19]. In cases where the participants are acquired through a crowdsourcing platform, the experimenter has little or no information regarding the background and profile of the crowdworkers. The absolute anonymity of subjects in an experiment is not often desirable. When specialized apparatus, hardware, software, or other equipment is required for a given experiment, leveraging crowdsourcing can be arduous, riddled with inconvenience, or in some cases even nearly impossible. Some ethnographic contexts in which crowdworkers participate in experiments may also be undesirable. These aspects, alongside

hidden confounding factors contribute to a lack of complete control over the subjects and the experimental environment.

1.3 Limitations of Existing Crowdsourcing Platforms for Academic Research

When considering using crowdsourcing for academic purposes, we must take into account the platform limitations. In particular, we must remember that these platforms, in general, were not built to support human-centered experiments, but rather for managing microtask units of work. The main purpose of most crowdsourcing systems is to provide a means to distribute the work and provide remuneration for it. As a result, researchers have described and created workarounds to help with these limitations.

A central limitation of using crowdsourcing platforms for human-centered experiments is ensuring that the participant is invested in the experimental tasks. This limitation is related to the absolute anonymity issue described above. Part of this limitation can be alleviated through using the participant reputation scores, but not entirely. As a result, experiments often employ a number of techniques. Consistency checks are conducted as a post process on the experimental results to ensure reasonably consistent answers for the same question or a set of sufficiently diverse answers [4,5,38]. Given drastically different answers for the exact same question (or the exact same answer for all questions even though they differ substantially), one could assume that the participants were not invested in the experimental tasks. Another method to ensure a high level of participant investment is to introduce special tasks in the experiment, or to use these special tasks as a pre-screening method for participants, to determine how much attention the participant is paying to the experiment [19,21]. Any combination of such techniques can be used to help ensure investments of the participants and the collection of high quality experimental data.

The above limitation is just one of many that we must consider when moving our experiments from the laboratory and deploying them in the crowd. Throughout this chapter, we bear in mind that crowdsourcing platforms were made to serve a different purpose and acknowledge the possible threats to validity in our experimental designs and deployments on crowdsourcing platforms.

2 Requirements for Human-Centered Experiments in the Laboratory

Having briefly discussed the limitations of conducting experiments in the laboratory, in this section we will elucidate the characteristics of human-centered experiments which are carried out in the laboratory. We address the goals of possible experiments which have an impact on the experimental structure, the resources needed for the setup, the participant pool, as well as the experimental process. We finish with a SWOT (*strengths, weaknesses, opportunities, threats*) analysis of laboratory experiments regarding these characteristics.

2.1 Goals of the Experiment

Human-centered experiments are no exception when it comes to requiring adequate planning to reach the preset objectives. Validity describes the degree to which the target has been reached, and is a key criterion for assessing the quality of the experiment. Other criteria are the reliability of the results, i.e. whether the results are stable when carrying out the experiment again (in terms of a parallel-test reliability, a re-test reliability, or an internal consistency within an experiment), the objectivity of the experiment, the economy of the process, the standardizability, the usefulness, and the comparability of results between experiments [6].

With respect to validity, generalizing the results of laboratory experiments carries with it the inherent disadvantage that the application of the research findings will normally be outside the laboratory. Thus, if the results themselves are to be applicable, laboratory experiments should be carefully designed to reflect a range of environmental, contextual and task characteristics of the situation in which they are to be applied. As an example, a laboratory experiment designed for finding out how well an object can be identified in an image (surveillance task) should be carried out using the same type of equipment (screen, ambient light situation, timing constraints) which will be used in the later surveillance situation. Otherwise, the experiment might be able to compare different experimental conditions well (relative validity), but not reflect the identification performance in an absolute way (absolute validity). On the other hand, in the case of laboratory experiments, the experimenter is in direct control of the environment. So even if the realistic use case cannot be fully simulated, it can be ensured that all participants in the experiment work with exactly the same hardware, under the same light and sound conditions, without external distractions and so on. Thus, confounding factors can be effectively reduced in a laboratory experiment.

As another example, an experiment might be designed in order to obtain an ordering of audiovisual stimuli which only differ to a very small extent. In such a case, this ordering might be better achieved in a laboratory than in a crowd environment, as the equipment used by the test participants can be controlled to a greater extent. It can be ensured that the test environment is mostly free of impediments (such as ambient noise or visual extractions) which would render the task more difficult, and thereby the test less sensitive for the given purpose.

The experimental situation also needs to be valid with respect to the involvement and potential collaboration of the participants. As an example, an experiment to analyze the communication quality of a Voice-over-IP system needs to be carried out in a conversational rather than a listening-only situation, because the Voice-over-IP system will mostly be used in a conversational mode. This can be reached quite easily in a laboratory situation by inviting test participants in pairs in order to carry out realistic conversations over the system, e.g. following pre-defined scenarios [29]. To do the same in a crowdsourcing environment would be far more difficult, as the scheduling of participants in the crowd is more difficult and might lead to timing and motivational conflicts.

Similarly, experiments designed to analyze usability and participant behavior in collaborative visualizations [32] may depend on direct interactions between participants. While distributed collaboration is often subject to the same kind of scheduling constraints as in the Voice-over-IP example above, experiments on co-located collaborative visualization are even harder to realize outside a laboratory environment. So in collaborative settings, the laboratory appears to have clear advantages.

Finally, some human-centered experiments require repeated participation in multiple phases of the experiment. For such experimental setups, it is crucial to have access to the same participants, maybe even groups of participants, after well-defined time intervals. In a laboratory experiment, participant selection according to these requirements is much easier to achieve than in currently available crowd platforms.

2.2 Resources

A major limitation of laboratory tests is the resources which are required in order to properly conduct an experiment. Formal laboratory tests require a considerable amount of time for the experimental planning, preparation of the environment, acquisition of suitable test participants, execution of the experiment, and finally analysis of the results, typically in the order of weeks or even months. Thus, a trade-off has to be made between the urgency with which the results of an experiment are needed, and the financial investment necessary to facilitate the laboratory experiment. The time which is necessary to carry out a formal laboratory experiment may also limit its applicability in iterative and agile product development cycles, which require iteration times of a week or less for each cycle in order to be efficient; a short timing may render laboratory tests incompatible with such development cycles.

Apart from the time, the test environment and the equipment which needs to be integrated into it are relevant resources. As mentioned above, the test environment is important to guarantee a high validity of the results, either in terms of ecological validity or in terms of sensitivity of the test procedure. Especially the latter requirement may cause high investments in terms of sound-insulated rooms (for sensitive auditory tests), rooms with controlled artificial lighting conditions (for visual tests), combinations of rooms with identical acoustic conditions (for conversation tests), and alike. It should be noted, however, that it is extremely difficult to achieve the same level of controlled and uniform environments in crowdsourced experiments. For highly sensitive experiments, the laboratory seems to be the best choice, despite the considerable investment costs. The investment to make a laboratory environment similar to a real-life usage scenario may be high: acoustic background noise may need to be inserted in a controlled but realistic way, dummy bystanders may need to be hired in order to simulate social presence, or additional furniture and accessories may be necessary to simulate a realistic atmosphere.

Integrated into the environment, the test equipment used by the participants may require further investments. In a laboratory experiment, it is easy to

guarantee that all participants use the same type of equipment (e.g. headphones, screens, interactive and connected devices) which has been controlled for its technical characteristics, and is monitored for proper functioning throughout the experiment. Such control is nearly impossible in a crowd-powered setting, where participants are expected to bring their own equipment, and where there is little or no control over the equipment. Having said that, alternate forms of crowdsourcing have been discussed in literature that overcome this issue, for example, by using public displays [22].

Finally, if the test equipment itself is part of the experiment [13], e.g., when testing immersive displays or virtual reality glasses, the required hardware may not even be freely available on the market or too expensive to expect at the disposal of crowdworkers. In many such situations again, there is no real alternative to running the experiments in a controlled laboratory setting or providing carefully selected test participants with the required hardware. Researchers have addressed this challenge by proposing methods to overcome equipment related obstacles in a few different domains [23,35].

2.3 Participant Pool

As the name suggests, human-centered experiments require human participants who act as "measuring organs". This renders such experiments "subjective", in the sense that human involvement is necessary to achieve the results, but they should be still "objective", such that the outcome is independent of the experimenter. However, the characteristics of the test participants will (and should) largely influence the test results.

According to the purpose of the experiment, participants can be classified according to their traits:

- perceptual and cognitive characteristics (hearing, vision, memory capacity, etc.)
- behavioral characteristics (left-handed vs. right-handed, dialect, sociolect, personality traits, etc.)
- experience and expertise (with the item under investigation, with similar items, with the domain, etc.)
- motivation (intrinsic or extrinsic motivation)
- individual preferences, capabilities or knowledge (sexual orientation, absolute hearing capability, individual background knowledge, language skills, etc.)
- personal characteristics (age, sex, level of education, nationality and cultural background, handicaps)

In a laboratory setting, participants may be selected and screened for all those characteristics which are deemed relevant for the outcome of the experiment. Unfortunately, this screening process is time-consuming, and may significantly limit the time available for the proper experiment. The availability of sufficient numbers of suitable participants with a particular set of characteristics may be very limited. In addition, in many cases the influence factors are not known

with respect to their (quantitative) impact, and it may be very difficult to find and access participants who show all relevant characteristics in a way which is representative for the actual use case (target user group). In such cases, one can assess the impact of participant characteristics on the test results only after the experiment. The result of this analysis may then limit the conclusions which can be drawn from the experiment.

The selection of test participants with desired characteristics is possible in a laboratory environment, albeit with a potentially high effort from the experimenter and significant compensation for the participants. For example, it may be possible to recruit computer-illiterate participants in order to test unbiased first-time usage of a computer system. This would be less probable for a crowd environment where participants are necessarily recruited through a computer platform, thus inherently limiting the pool of test participants to those with certain characteristics. To overcome this issue, some platforms offer an API to select workers with certain desired characteristics. For example, CrowdFlower[2] offers three levels of crowdworkers based on their reputation and quality of work.

2.4 Process and Control

In a laboratory setting, the experimental process can be properly designed and closely controlled to achieve an optimum reliability of the results in terms of accuracy and validity. For example, test participants can be properly screened with respect to all their relevant characteristics, and the screening process can be adequately supervised to guarantee that no cheating is possible. In addition, participants can be instructed in a standardized way, giving room for individual questions they might have in order to ascertain their complete understanding of the experimental task at hand. The design and timing of individual tasks and sessions can be closely controlled in order to limit fatigue or mental overload. In addition, the motivation of the test participants can be better controlled, so as to avoid participants "mechanically" resolving the given tasks without making use of the human capabilities which are at the core of the experiment. The simple presence of a human experimenter in the test laboratory, and the social facilitation of talking to him/her and receiving the instructions in a personalized way, may increase the reliability of the results. In addition, participants can easily access the experimenter in case questions or problems arise during the test run.

If the experimental design requires to split tasks across multiple sessions, the experimenter can recruit the same participants again for multiple sessions, thus facilitating a within-subject design. Such designs are more difficult to achieve in a crowd setting, where tasks are usually small and short in duration, and where extrinsic motivation is a big factor that affects participation.

[2] http://crowdflower.com/ last accessed 14 Jun 2017.

2.5 SWOT Analysis of Human-Centered Laboratory Experiments

In the following table, we analyze and present the strengths, weaknesses, opportunities and threats that entail the running of human-centered experiments in laboratories.

STRENGTHS	WEAKNESSES
• high level of control over experimental process and environment • reliability of participants • participant screening for special skills and characteristics	• limited participant pool • time-consuming • expensive • artificial, simulated environment
OPPORTUNITIES	**THREATS**
• collaborative experiments • multi-phase experiments • personal interaction and feedback channels • use of specialized hardware	• limited ecological validity • draw conclusions which may not hold in real life

3 Transition to Using Crowdsourcing for Human-Centered Experiments

In this section we discuss how the different dimensions of a human-centered experiment can be carried out using crowdsourcing. We analyze how characteristic features of crowdsourcing can be exploited in order to run human-centered experiments using the crowd.

3.1 Goals of the Experiment

Crowdsourcing tasks can be executed with a variety of goals, ranging from generating data to building ground truths for evaluation. Previous work has categorized typical crowdsourcing microtasks into an exhaustive taxonomy at the top-level based on the goals of a task requester or experimenter [18]. These categories were determined to be: *information finding*, *verification and validation*, *content creation*, *interpretation and analysis*, *surveys*, and *content access*. Most

commonly crowdsourcing has been used as a tool to solve problems that require human intelligence or input at scale. However over the last few years, researchers have begun considering the paid crowdsourcing paradigm as a potential avenue to run scientific experiments that were previously conducted and constrained in laboratory settings [7,26,37,41]. When it comes to the validity of conducting a human-centered experiment using crowdworkers, the ease with which a diverse and representative population can be acquired is a big advantage. Through the course of this section, we will explore the inherent characteristics of crowdsourcing that need to be further considered to run valid human-centered experiments in the crowd.

3.1.1 Collaboration Between Participants

In a standard microtask crowdsourcing scenario each worker typically contributes independently to the final result. Nevertheless, if an experiment needs the collaboration between subjects, the crowdsourcing scenario can be adapted accordingly. 'Games with a purpose' are a good example of such collaboration, where people collaborate in order to solve different problems, ranging from image tagging [49] to identification of protein structures [33]. Recent work has also shown that team competition designs can be effective in improving the throughput of crowdsourced tasks [46].

On the other hand, none of the primary microtask crowdsourcing platforms (such as Amazon's Mechanical Turk (AMT)[3] or CrowdFlower[4]) facilitate direct collaboration between workers, so the coordination between subjects must be manually implemented and facilitated externally. Furthermore, imposing a schedule and time constraints on the workers may hurt their spirits and increase dropouts. For instance, when proper collaboration means are not employed, a worker may either have to wait for long periods of time before his collaborators are found, or he could be paired with a low quality or undesirable workers.

3.1.2 Multi-phase Experiments with the Same Set of Participants

In case of experiments composed by different repeated phases, where a fundamental requirement is to involve the same set of participants in each phase, the anonymity of the subjects characterizing the crowdsourcing environment makes the execution of such types of experiments very challenging, since the only possibility is to directly contact the worker (typically via email). Hence, if a crowdsourcing platform does not disclose contact information or it does not facilitate reaching particular workers directly, a possible solution is to redirect workers to a customized external platform, where the information needed can be collected in order to contact the same subjects in future. In prior work, authors proposed a two-stage implementation of crowdsourcing for QoE assessment [27].

[3] https://www.mturk.com/ last accessed 14 Jun 2017.
[4] http://www.crowdflower.com/ last accessed 14 Jun 2017.

Although freelancing or expert-sourcing platforms such as Upwork[5] facilitate collaboration between participants to complete complex tasks in multiple phases if required, they are less-suitable for human-centered experiments, and beyond the scope of this work.

3.2 Resources

The main characteristic of an experiment performed with the crowd is that each subject uses his own device. As a consequence the time required for environment preparation is curtailed to a large extent; there is no need to prepare the laboratory or to configure the equipment. At the same time, an experimenter has no direct control over the hardware and software configuration with respect to the subjects' environments. This may be particularly detrimental if the experiment requires special hardware, or specific software configurations to ensure validity of the results. It is cumbersome to impose any type of control on the environment with the aim to either create a uniform setting across participants, or to make it more similar to the real-life usage scenario. However, it is still possible to check the reliability of the worker hardware and software using scripts that run on the worker's device reporting its configuration in term of browser version, operative system, hardware configuration and so forth. With this information it is possible to pre-screen the workers who don't satisfy the minimal requirements needed for the experiment.

The cost of setting up the experiment in terms of equipment is virtually zero, but we need to take into the account the costs in terms of effort in designing the crowdsourcing task so as to satisfy the requirements of the experiment. This cost increases exponentially if a specific feature needs to be completely implemented from scratch, due to a lack of support on the crowdsourcing platform of choice. A larger effort is required to implement software compatible to various web browsers, supporting various devices, and so forth. Further, (offline) processing of results requires extra efforts and the monitoring of hidden influence factors needs to be implemented in the test design; all accounting for additional costs. In addition, if the paid crowdsourcing paradigm is employed, then participants need to be monetarily compensated.

3.3 Participant Pool

Some of the key implications of crowdsourcing human-centered experiments with respect to the participant pool, arise from the inherent characteristics of the paradigm, and are presented below.

- *Quantity*: An experimenter can attain access to an extremely large population size via various crowdsourcing platforms. Thus, laboratory experiments which were previously constrained to the order of tens or hundreds of experiment subjects can scale-up to the order of thousands of participants without huge ramifications on the costs entailed.

[5] https://www.upwork.com/ last accessed 14 Jun 2017.

- *Availability*: Laboratory experiments are typically constrained by the availability of subjects, as well as open hours of the laboratory itself. The transition of such experiments to using crowdsourcing would mean that participants would be available around the clock, and the experimenter would not necessarily be restricted by the time of the day.
- *Diversity & Reachability*: Crowdworkers that can be reached via crowdsourcing platforms constitute a highly diverse population, covering a wide range of demographic attributes (age, gender, ethnicity, location, and so forth). Thus, a human-centered experiment can benefit from this diversity and consequently arrive at more representative results.
- *Quality & Reliability*: One of the major challenges in exploiting the prowess of crowdsourcing for human-centered experiments is quality control and the reliability of participants. Experiments conducted in a laboratory can benefit from surveillance of the subjects, thereby eliciting adequate behavior and ensuring reliable participation. Over the last few years, researchers have devised a number of quality control mechanisms in crowdsourcing ranging from task design methods, to worker pre-selection, or even post-hoc analysis [11,19,36]. Therefore, although there are additional costs entailed to sustaining the reliability of participants in crowdsourced human-centered experiments, it is certainly possible to achieve.

3.4 Process and Control

A number of aspects need to be considered in order to exercise control over human-centered experiments when using crowdsourcing.

- *Design*: Additional effort is required to design an experiment that is suitable for the participation of crowdworkers. The use of standard microtask crowdsourcing platforms as a source of acquiring subjects for human-centered experiments, means that the experiments may have to be decomposed into micro units of work.
- *Incentives*: A variety of incentives have been used to encourage participation in laboratory experiments previously, such as course credits, monetary compensations, altruistic intent, and so forth. When microtask crowdsourcing platforms are employed for human-centered experiments, the typical mode of participant acquisition is through financial incentives. The entailing costs depend on the complexity of the experiment, the effort required from participants, and amount of time required for task completion.
- *Personal Touch, Social Facilitation, & Feedback Channels*: One of the limiting factors in crowdsourcing human-centered experiments is the lack of personal interaction between the experimenter and the participants. Experimenters benefit in laboratories from facilitating the subjects and providing them with immediate feedback where required. Microtask crowdsourcing platforms typically provide feedback channels with limited flexibility (for example, via chat

rooms or emails). Thus, additional efforts are required from the experimenter to ensure that participants are adequately facilitated and have understood their task objectives sufficiently [27]. Unlike in laboratory environments, subjects cannot be monitored easily and there is lesser control over the experimental protocol.

- *Equipment Configuration*: Human-centered experiments which require specific equipment or special devices (for example, ECG machines), or those that require participants to be embedded in the same environments (screen-resolution, distance to the screen, ethnographic contexts, software/hardware configurations, and so forth), are less suitable for the transition to using crowdsourcing. Although there are ways to pre-select crowdworkers in order to satisfy the requirements, this requires additional effort.
- *Optimization*: A big advantage of running human-centered experiments using crowdsourcing is the potential to optimize for given needs (such as accuracy of crowdworkers, or the amount of time within which responses are to be gathered). If the most important criteria of the experiment is to ensure reliable responses from every participant, then one can leverage the in-built filters on the crowdsourcing platform, apart from exercising additional external guidelines [19]. This may lead to longer task completion times. However, if time is of essence then one can assume a more liberal means of allowing participation, and thereafter employ post-hoc analysis to filter out undesirable subjects. The scalability of crowdsourcing allows for such optimization as per the requirements at hand.

3.5 SWOT Analysis for Crowdsourced Human Experiments

Previous works have discussed the role of crowdsourcing in human experiments [44]. Horton et al. showed that experiments using crowdsourcing are valid internally and can be valid externally, just as laboratory experiments [26]. Similarly, Crump et al. evaluated the use of Amazon's Mechanical Turk to conduct behavioral experiments by replicating a variety of tasks from experimental psychology [7]. The authors found that most of the replications were successful, while a few exhibited a disparity with respect to laboratory results. They assert that despite the lack of environmental control while using crowdsourcing, the standardization and control over experiment procedures is an advantage.

We analyze the strengths, weaknesses, opportunities and threats that entail running human-centered experiments using crowdsourcing in the following table.

STRENGTHS	WEAKNESSES
• Ease of access to diverse and representative populations • Large-scale experiments are feasible • Time-efficiency • Flexibility with time of the day, duration of experiments • Relatively inexpensive	• Less control over the experimental environment • Extra effort required for collaborative or multi-phase experiments • Lack of knowledge regarding participants' background
OPPORTUNITIES	**THREATS**
• Optimization of experiment configuration (time, quality, and reliability) • New possibilities to broaden the research in various domains. For example collaboration and interaction between users, real-life environment (heterogeneous client devices and software, various network access technologies).	• Limited absolute validity of experiment results • Additional technical constraints such as bandwidth, client device compatibility, web-based frameworks, contextual monitoring, etc.

4 Methodological Considerations

As observed through the course of this chapter, using the crowd for performing human-centered experiments provides different opportunities but also raises several challenges. In this section, we discuss existing solutions and propose new approaches to address the concomitant challenges.

4.1 Challenges and Opportunities

Crowdsourcing creates several opportunities for performing human-centered experiments. It provides a fast way to access a wide set of participants, it does not require set up time and it allows to optimize the configuration of an experiment.

4.1.1 Existing Platforms Demand Workarounds – Current Solutions

We note that existing microtask crowdsourcing platforms are not directly meant for human-centered experiments. While platforms for academic research are on the rise (as pointed out in Chap. 4), they are not yet sufficiently established to suit global needs. However, to overcome shortcomings of existing platforms, several workarounds have been proposed over the last decade that address many challenges. We elaborate on the key features of crowdsourcing microtasks that have attracted adequate solutions.

- *Quality Control.* Due to the lack of direct control and supervision over participants in crowdsourced tasks, quality control has been identified as a pivotal aspect that determines the effectiveness of the paradigm. Many mechanisms have been proposed to assert the quality of results produced through crowdsourced tasks. Proposed solutions include the use of gold-standard questions [9,11,40], attention check questions, consistency checks, and psychometric methods [36], worker behavioral metrics and optimal task design [19], feedback and training [10,17], and optimizing task parameters such as task length and monetary compensation [3,20,37]. Qualification tests and pre-screening methods have also been adopted in order to select appropriate workers for a given task. These existing quality control mechanisms can be easily applied when running human-centered experiments using the crowd.
- *Improving Effectiveness.* Several optimization techniques have been introduced in prior works in order to increase the throughput of crowdworkers, maximize the cost-benefit ratio of deploying crowdsourced microtasks [45,46], and improving the overall effectiveness of the microtask crowdsourcing model. Gamification has been shown to improve worker retention and throughput of tasks [12]. Other works have suggested pricing schemes, or achievement priming to retain workers and improve latency in crowdsourced microtasks [8,16]. Similar strategies can be adopted where applicable, while running human-centered experiments using the crowd.

4.1.2 Elegant Solutions – An Outlook for Future Crowdsourcing Platforms

Owing to the great opportunities that crowdsourcing provides for human-centered experiments that were priorly constrained to the laboratory, we envisage a future where crowdsourcing platforms directly support and facilitate greater control to run human-centered experiments in the crowd.

- *Tailored Platforms.* First and foremost, there is a need for tailored platforms that support human-centered experiments. Due to the fact that traditional microtask crowdsourcing platforms have not been built to facilitate human-centered experiments in particular, workarounds are required to execute such experiments using these platforms. Some steps have already been taken towards building such tailored solutions; a good example is that of GraphUnit, a framework for visualization evaluation that leverages crowdsourcing [39].
- *Feedback & Supervision.* Experiment and task administrators currently use implicit feedback channels such as emails or chat rooms to communicate with crowdworkers. Enabling real-time interaction between crowdworkers and the task administrators can go a long way towards the social facilitation of potential experiment subjects in the human-centered experiments.
- *Iterative Design.* Human-centered experiments may require to be carried out in multiple phases using the same set of participants. Thus, platforms need to accommodate such iterative designs of experiments.

- *Worker Profiles.* Elaborate worker profiles that include the skills and interests of crowdworkers (similar to freelancing platforms), and their demographic details need to be made available to the task administrators. Such transparency will enable a seamless match-making process between available experiments and suitable crowdworkers on the platform. See Chap. 3 for a detailed discussion on worker profiles.

4.1.3 Task Complexity

In behavioral research and psychology, the impact of task complexity in various domains has been studied well [2]. Similarly, in the microtask crowdsourcing paradigm, *task complexity* is a complicated aspect that depends on several factors. There has been little research that deliberates on the impact of task complexity on various aspects of crowdsourcing such as worker performance, worker retention rates, and motivation. In order to create crowdsourcing solutions that are generalizable across different types of tasks, we need to consider the aspect of task complexity. Jie et al. recently showed that task complexity is perceived coherently among crowdworkers, and that it is effected by the type of the task [52]. The authors proposed several structural features to model and measure task complexity. We highlight the consideration of task complexity as an important opportunity for future research.

4.2 Guidelines and Ethics: How Do Ethical Values Transfer to Crowdsourced Human Experiments?

The major ethical concerns with microtask crowdsourcing platforms yield from the fact that a considerable number of workers contributing on these platforms earn their livelihood from this work [30,31]. Hence, workers need to be adequately compensated in accordance to the time and effort exerted through their contribution to crowdsourced tasks. A variety of aspects such as task pricing, clarity [15], complexity, and so forth affect crowd work and need to be considered to ensure fair and healthy dynamics between the workers and requesters. The manual labor of crowdworkers was further recognized in recent times by the sentence against CrowdFlower, which undercut the United States minimum wage legislation [50].

We list a few ethical concerns arising from current practice in microtask crowdsourcing platforms. For a more elaborate discourse on ethical values in crowdsourcing human experiments, see Chap. 3.

- Lack of adequate communication channels between workers and task requesters or experimenters. Thus, crowdworkers cannot appeal against declined work or take corrective measures when tasks are misunderstood.
- No guarantee for payments promised as compensation, the task requester has all the power to credit or discredit contributions from crowdworkers.

- Monetary compensation in return for crowd work does not always meet the minimum wage stipulations.
- Often studies on crowdsourcing platforms do not go through ethical review boards of research institutions.

According to [51] it is not sufficient from an ethics point of view to voluntarily increase the rate of payment for Amazon's Mechanical Turk (AMT) tasks as it won't resolve the fundamental inequities of the precarious employment situation of a considerable number of workers. Recent works have addressed the concerns yielding from the power asymmetry in crowdsourcing microtask workflows, with an aim to pave a way towards an ethically balanced paradigm of crowd work [47]. Guidelines to practice ethical crowdsourcing as task requesters from a holistic standpoint have been defined in previous work [28].

5 Future of Crowdsourcing Human Experiments

In this chapter we have discussed and elucidated the opportunities of running human-centered experiments in the crowd. We note that the crowdsourcing paradigm provides a unique means to scale up otherwise constrained laboratory experiments. Although there are a few disadvantages of running human-centered experiments in the crowd as noted earlier, the benefits of using crowdsourcing outweigh the threats in the applicable scenarios.

5.1 Crowdsourcing and Laboratory Experiments - A Complimentary Perspective

In the end it is unlikely that crowdsourcing will replace lab testing altogether. A more likely scenario is that experimenters will learn how best to combine crowd and lab to balance the benefits and drawbacks of each. These mixed-method investigations hold a great deal of promise for creating models that are both highly predictive and generalizable to diverse populations of interest. We will discuss a few examples of ways in which this might be done in the hope that it may inspire new and better approaches to human experimentation.

5.1.1 Lab First, Crowd Second: Evaluation of Theories Generated from Laboratory Studies

While it is tempting for interface designers to directly apply the results of an experiment to an interface design, we must keep in mind that many of these studies were designed to contribute to a natural science of human cognition. Accordingly, the phenomena they describe are not intended to be directly applied to an interface but are instead a means to the end of generating and testing theories of human information processing that are applicable to a broad range of situations. Taking Pylyshyn's FINST theory [43] as an example we can see how studies conducted with a variety of tasks and stimuli including multiple object tracking, subitizing, and visual search were designed specifically to test whether

our visual system had a finite number of visuospatial attentional tokens that could facilitate performance of a variety of tasks. These generalizable theories are considered architectural in that they provide specific capabilities that can be assembled in different ways to accomplish different tasks. A key aspect of the research agenda in cognitive science is to identify these capabilities and to assemble them in the form of an overall cognitive architecture, such as Anderson's ACT-R [1]. Indeed, while many in the visualization and HCI communities are aware of Pirolli and Card's Sensemaking theory [42], few are aware that one of the goals of this work was to facilitate application of ACT to sensemaking in Fu and Pirolli's SNIF-ACT model [14].

Because of the need for control of the experimental situation and exploration of the parameter space of these models it is hard to imagine that theory at the level of cognitive architecture could be generated using crowdsourcing methods. Where crowdsourcing might play a role would be in evaluating these models in the context of the more diverse set of participants and situations of use. The research question here would be whether those models can be parameterized in such a way that they can account for a diversity of people and situations.

5.1.2 Crowd First, Lab Second: Identifying Key Individuals and Sub-populations for Future Studies

Many of the more compelling studies in cognitive neuroscience are conducted with the participation of those rare individuals who differ from the general population. Whether it is due to genetics, a neurological accident, or an unusual training experience these extreme cases can give us insight into human limitations and capabilities. One crowdsourcing example comes from Philip Tetlock and Barbara Meller's Good Judgment Project [48]. In this project the researchers crowdsourced predictions about a variety of political developments from over 2000 participants in order to identify a sub-population of individuals who were consistently accurate over time. These individuals were then tested to determine how they differed from the general population. Bringing these individuals into controlled testing situation might well prove effective in establishing more robust cognitive architectures and assessing the range of operating parameters that can be found in the overall population.

5.2 Conclusions

We are only beginning to understand how to best utilize crowdsourcing for human-centered experimentation. The ease with which a large number of participants having desirable traits can be found, the scalability of experiments, the efficiency with respect to time and entailing costs, the flexibility with the time of the day and duration of experiments, makes the crowdsourcing of human-centered experiments very promising. Challenges that pertain to the lack of control over the experimental environment can be overcome to an extent, through prudent experimental design choices and manipulating crowdsourcing task workflows to suit requirements. As we continue to explore the optimum trade-offs between the

laboratory and the crowd, we will discover new ways to manage task allocation and delivery, coordination of multiple crowdworkers in collaborative and competitive task performance, and new data analysis methods that can be brought to bear on the rich datasets that can be produced with crowd and mixed method experimentation.

Acknowledgment. We would like to thank Dagstuhl for facilitating the seminar (titled, '*Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments*') that brought about this collaboration. Part of this work (Sect. 4) was supported by the German Research Foundation (DFG) within project A05 of SFB/Transregio 161. We also thank Andrea Mauri and Christian Keimel for their valuable contributions and feedback during discussions.

References

1. Anderson, J.R., Matessa, M., Lebiere, C.: ACT-R: a theory of higher level cognition and its relation to visual attention. Hum. Comput. Interact. **12**(4), 439–462 (1997)
2. Campbell, D.J.: Task complexity: a review and analysis. Acad. Manag. Rev. **13**(1), 40–52 (1988)
3. Cheng, J., Teevan, J., Bernstein, M.S.: Measuring crowdsourcing effort with error-time curves. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1365–1374. ACM (2015)
4. Chung, D.H.S., Archambault, D., Borgo, R., Edwards, D.J., Laramee, R.S., Chen, M.: How ordered is it? On the perceptual orderability of visual channels. Comput. Graph. Forum **35**(3), 131–140 (2016). (Proc. of EuroVis 2016)
5. Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., Singh, M.: How well do line drawings depict shape? ACM Trans. Graph. **28**(3), 1–9 (2009)
6. Cozby, P.: Asking people about themselves: survey research. In: Methods in Behavioral Research, 7th edn., pp. 103–124. Mayfield Publishing Company, Mountain View (2001)
7. Crump, M.J., McDonnell, J.V., Gureckis, T.M.: Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PloS one **8**(3), e57410 (2013)
8. Difallah, D.E., Catasta, M., Demartini, G., Cudré-Mauroux, P.: Scaling-up the crowd: micro-task pricing schemes for worker retention and latency improvement. In: Second AAAI Conference on Human Computation and Crowdsourcing (2014)
9. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Mechanical cheat: spamming schemes and adversarial techniques on crowdsourcing platforms. In: CrowdSearch, pp. 26–30. Citeseer (2012)
10. Dow, S., Kulkarni, A., Klemmer, S., Hartmann, B.: Shepherding the crowd yields better work. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 1013–1022. ACM (2012)
11. Eickhoff, C., de Vries, A.P.: Increasing cheat robustness of crowdsourcing tasks. Inf. Retr. **16**(2), 121–137 (2013)
12. Feyisetan, O., Luczak-Roesch, M., Simperl, E., Tinati, R., Shadbolt, N.: Towards hybrid NER: a study of content and crowdsourcing-related performance factors. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9088, pp. 525–540. Springer, Cham (2015). doi:10.1007/978-3-319-18818-8_32

13. Fikkert, W., D'Ambros, M., Bierz, T., Jankun-Kelly, T.J.: Interacting with visualizations. In: Kerren, A., Ebert, A., Meyer, J. (eds.) Human-Centered Visualization Environments. LNCS, vol. 4417, pp. 77–162. Springer, Heidelberg (2007). doi:10.1007/978-3-540-71949-6_3
14. Fu, W.T., Pirolli, P.: SNIF-ACT: a cognitive model of user navigation on the world wide web. Hum. Comput. Interact. **22**(4), 355–412 (2007)
15. Gadiraju, U.: Crystal clear or very vague? Effects of task clarity in the microtask crowdsourcing ecosystem. In: 1st International Workshop on Weaving Relations of Trust in Crowd Work: Transparency and Reputation Across Platforms, Co-located With the 8th International ACM Web Science Conference 2016, Hannover (2016)
16. Gadiraju, U., Dietze, S.: Improving learning through achievement priming in crowdsourced information finding microtasks. In: Proceedings of ACM LAK Conference. ACM (2017, to appear)
17. Gadiraju, U., Fetahu, B., Kawase, R.: Training workers for improving performance in crowdsourcing microtasks. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 100–114. Springer, Cham (2015). doi:10.1007/978-3-319-24258-3_8
18. Gadiraju, U., Kawase, R., Dietze, S.: A taxonomy of microtasks on the web. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, pp. 218–223. ACM (2014)
19. Gadiraju, U., Kawase, R., Dietze, S., Demartini, G.: Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015), Seoul, 18–23 April 2015, pp. 1631–1640 (2015)
20. Gadiraju, U., Siehndel, P., Fetahu, B., Kawase, R.: Breaking bad: understanding behavior of crowd workers in categorization microtasks. In: Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 33–38. ACM (2015)
21. Gardlo, B., Egger, S., Seufert, M., Schatz, R.: Crowdsourcing 2.0: enhancing execution speed and reliability of web-based QoE testing. In: Proceedings of the IEEE International Conference on Communications (ICC), pp. 1070–1075 (2014)
22. Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., Kostakos, V.: Crowdsourcing on the spot: altruistic use of public displays, feasibility, performance, and behaviours. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 753–762. ACM (2013)
23. Hanhart, P., Korshunov, P., Ebrahimi, T.: Crowd-based quality assessment of multiview video plus depth coding. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 743–747. IEEE (2014)
24. Heer, J., Bostock, M.: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010), Atlanta, 10–15 April 2010, pp. 203–212 (2010)
25. Heinzelman, J., Waters, C.: Crowdsourcing crisis information in disaster-affected Haiti. US Institute of Peace (2010)
26. Horton, J.J., Rand, D.G., Zeckhauser, R.J.: The online laboratory: conducting experiments in a real labor market. Exp. Econ. **14**(3), 399–425 (2011)
27. Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans. Multimed. **16**(2), 541–558 (2014)
28. Hoßfeld, T., Tran-Gia, P., Vucovic, M.: Crowdsourcing: from theory to practice and long-term perspectives (Dagstuhl Seminar 13361). Dagstuhl Rep. **3**(9), 1–33 (2013). http://drops.dagstuhl.de/opus/volltexte/2013/4354

29. ITU-T Rec. P.805: Subjective evaluation of conversational quality. International Telecommunication Union, Geneva (2007)
30. Ipeirotis, P.G.: Analyzing the Amazon Mechanical Turk marketplace. XRDS: Crossroads ACM Mag. Stud. **17**(2), 16–21 (2010)
31. Ipeirotis, P.G.: Demographics of Mechanical Turk (2010)
32. Isenberg, P., Elmqvist, N., Scholtz, J., Cernea, D., Ma, K.L., Hagen, H.: Collaborative visualization: definition, challenges, and research agenda. Inf. Vis. **10**(4), 310–326 (2011)
33. Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popović, Z., Baker, D., Players, F.: Algorithm discovery by protein folding game players. Proc. Natl. Acad. Sci. **108**(47), 18949–18953 (2011)
34. Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al.: Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat. Struct. Mol. Biol. **18**(10), 1175–1177 (2011)
35. Lebreton, P.R., Mäki, T., Skodras, E., Hupont, I., Hirth, M.: Bridging the gap between eye tracking and crowdsourcing. In: Human Vision and Electronic Imaging XX, San Francisco, 9–12 February 2015, p. 93940W (2015)
36. Marshall, C.C., Shipman, F.M.: Experiences surveying the crowd: reflections on methods, participation, and reliability. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 234–243. ACM (2013)
37. Mason, W., Suri, S.: Conducting behavioral research on Amazons Mechanical Turk. Behav. Res. Methods **44**(1), 1–23 (2012)
38. McCrae, J., Mitra, N.J., Singh, K.: Surface perception of planar abstractions. ACM Trans. Appl. Percept. **10**(3), 14: 1–14: 20 (2013)
39. Okoe, M., Jianu, R.: GraphUnit: evaluating interactive graph visualizations using crowdsourcing. Comput. Graph. Forum **34**(3), 451–460 (2015)
40. Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., Biewald, L.: Programmatic gold: targeted and scalable quality assurance in crowdsourcing. In: Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence (WS-11-11). AAAI (2011)
41. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on Amazon Mechanical Turk. Judgm. Decis. Mak. **5**(5), 411–419 (2010)
42. Pirolli, P., Card, S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: Proceedings of International Conference on Intelligence Analysis, vol. 5, pp. 2–4 (2005)
43. Pylyshyn, Z.W.: Things and Places: How the Mind Connects with the World. MIT Press, Cambridge (2007)
44. Rand, D.G.: The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. J. Theor. Biol. **299**, 172–179 (2012)
45. Rokicki, M., Chelaru, S., Zerr, S., Siersdorfer, S.: Competitive game designs for improving the cost effectiveness of crowdsourcing. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 1469–1478. ACM (2014)
46. Rokicki, M., Zerr, S., Siersdorfer, S.: Groupsourcing: team competition designs for crowdsourcing. In: Proceedings of the 24th International Conference on World Wide Web, pp. 906–915. International World Wide Web Conferences Steering Committee (2015)

47. Salehi, N., Irani, L.C., Bernstein, M.S., Alkhatib, A., Ogbe, E., Milland, K., et al.: We are dynamo: overcoming stalling and friction in collective action for crowd workers. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1621–1630. ACM (2015)
48. Tetlock, P.E., Mellers, B.A., Rohrbaugh, N., Chen, E.: Forecasting tournaments tools for increasing transparency and improving the quality of debate. Curr. Dir. Psychol. Sci. **23**(4), 290–295 (2014)
49. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326. ACM (2004)
50. Weber, L., Silverman, R.E.: On-demand workers: we are not robots. Wall Str. J. 7 (2015)
51. Williamson, V.: On the ethics of crowdsourced research. PS Political Sci. Politics **49**(01), 77–81 (2016)
52. Yang, J., Redi, J., DeMartini, G., Bozzon, A.: Modeling task complexity in crowdsourcing. In: Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2016), pp. 249–258. AAAI (2016)