Online Playtesting With Crowdsourcing: Advantages and Challenges

Florian Brühlmann
Center for Cognitive
Psychology and Methodology
University of Basel
4055 Basel, Switzerland
florian.bruehlmann@unibas.ch

Elisa D. Mekler
Center for Cognitive
Psychology and Methodology
University of Basel
4055 Basel, Switzerland
elisa.mekler@unibas.ch

Gian-Marco Schmid
Center for Cognitive
Psychology and Methodology
University of Basel
4055 Basel, Switzerland
gian-marco.schmid@unibas.ch

Abstract

Answering important design questions and delivering actionable insights within a couple of days is invaluable. Traditional playtests are often time consuming, expensive and deliver insights based on only a small sample of participants. Crowdsourced playtests may deliver comparable quality of feedback with less resources. However, several aspects have to be considered in order to receive meaningful and actionable results. Based on our experience, we provide five recommendations to ensure data quality and prevent fraud. Taken together, this suggests that crowdsourced playtesting is a promising alternative for indie, non-profit and academic Games User Research.

Author Keywords

Games User Research; Playtesting; Crowdsourcing; Experiments

ACM Classification Keywords

H.5.2 [User Interfaces]: Evaluation/methodology; K.8.0 [Personal Computing]: Games

Introduction

Playtests are one of the most important practices in Games User Research (GUR) and can to a large degree influence whether the final product will be commercially viable. For these tests a researcher usually invites individuals from the target audience to play a game (or parts of a game) in development to identify design flaws and gather feedback. However, conducting playtests is expensive because one has to acquire participants, observers and the necessary equipment. Additionally, conducting these studies can take a lot of time, depending on the availability of participants from the target audience, the elaborateness of the used methods and the required sample size for statistical analysis. Online evaluations are already common practice for digital games. For instance, online beta tests or early-access phases ask members of the target audience to evaluate the game. However, in order to conduct these online beta tests developers need the necessary infrastructure and a sizable community. As resources, both in money and in time, are sparse in independent, non-profit or academic game development projects, this option is often not possible on a large scale. Additionally, playtesting should begin before reaching the beta phase, ideally as early as possible, to reduce blind spots of development.

A promising alternative are crowdsourcing platforms, for example CrowdFlower or Amazon's Mechanical Turk, on which users can complete small tasks for a (usually monetary) incentive [4]. These services have the advantage of a large force of readily available users with a variety of backgrounds and have been found to be reliable for behavioral research and user studies [5]. Crowdsourced playtesting is low-priced and therefore a promising avenue for small and indie game developers. People generally like to engage in crowdsourced micro-tasks and the monetary reward is not the main motive for participation [1]. In this position paper, we describe our experience in conducting crowdsourced playtests and report recommendations that are based on what we learned. Our experience comes mainly from two projects; the development and evaluation of a puzzle game and a study that asked participants to play and compare two existing indie games (Canabalt and Super Hot, refer to [2]). Our goal is to encourage research on and practice of crowdsourced GUR.

Typical procedure

Crowdsourced playtests can be structured in the same way as an online survey. The required infrastructure is often already existing or can be acquired at low cost (e.g., a simple web server to host an online survey application). Usually, a description of the task, the necessary inclusion criteria for participation and the expected reward are displayed on the crowdsourcing platform, together with a link to website where the survey and the game is hosted. Upon completion of the survey and after answering any attention check items correctly, participants are given a completion code that can be redeemed for the reward on the crowdsourcing platform.

Advantages

The main advantages of crowdsourced online playtests are their low cost and their prompt results. In our experience with CrowdFlower, collecting 50 responses of a 30 minutes study takes often no longer than half a day. This reduces the data collection time significantly. We found that a compensation of $1.50 for a 30 minutes study works well in terms of data quality, completion time and satisfaction of the participants (i.e., high feedback ratings). Thus, conducting 50 playtests costs no more than $100 (including service fees), making it very cost efficient. In terms of data quality and incentive, paying only a small amount for the completion of the survey and giving participants who responded carefully a bigger bonus has in our experience increased data quality and prevented malicious behavior.

Another advantage is the availability of a broad population of participants with a variety of demographic characteristics and interests. For Amazons' crowdsourcing service

Mechanical Turk, research suggest that the population of workers is demographically more diverse than standard internet samples or typical college samples and the data is as good or better and as reliable as those obtained with traditional methods [3]. It is also possible to specifically recruit individuals that are not as easily reachable in some game communities such as older or female players. This helps to reach specific target groups and to reduce possibly biased feedback from participants that are fans. However, because of this diverse population there is the risk that some participants might not like the theme of the game. In our study, participants who disliked violence in digital games rated the game Super Hot very low on player experience scales. We were only able to detect this issue (and a few technical issues) because we gave people the option comment. Therefore, it is beneficial to provide users the ability to comment on their response to an item or to specifically ask for a reason why they give a low or high rating.

The scalability of crowdsourced playtests is another advantage; in one of our projects we were able to identify issues in the difficulty adaptiveness because of 35 players who played a variant of the game via CrowdFlower. Figure 1 depicts that the adaptive difficulty was working, however, the game was too difficult in the first round and almost all participants lost many rounds before winning once. We would not have been able to spot this so easily with a small sample of colleagues in a hallway usability testing.

However, conducting research online and especially on crowdsourcing platforms requires a well tested set-up and a few measures to ensure data quality.

Challenges

One of the key limitations of crowdsourcing playtests is that the game usually needs to run in a web browser. While this is a serious limitation for games with highly realistic graphics, many independent or non-profit games do not rely on fancy graphics and expensive game engines. Aside from Flash and JavaScript based games, games developed with Unity can be exported for the Unity Webplayer and run in any modern browser. Depending on the policy of the crowdsourcing service, it might also be possible to let people download the game, but retaining control over the game and the procedure of the playtest might be more difficult.

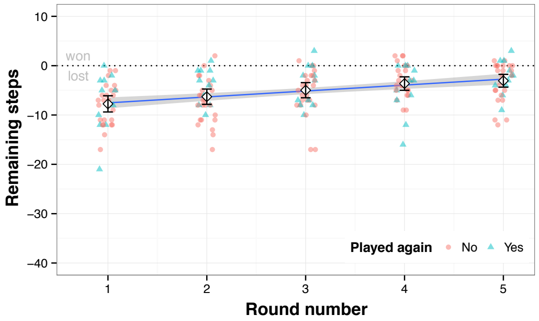

Figure 1: An example of behavioral data from a playtest. Participants played 5 rounds of a game where they had to solve a puzzle with less than 25 steps to win the round.

Another limitation is the duration of the study. To our knowledge most tasks or "survey jobs" on crowdsourcing platforms take only a couple of minutes up to half an hour. While this might be sufficient to test effects of small changes in mechanics or other game elements, this timeframe might be too limited to test more complex parts (e.g., credibility of the story). There is currently only little knowledge about the feasibility of longer online playtesting sessions.

It can be expected that responses from participants recruited by traditional means to be comparable to those from crowdsourcing platforms (e.g., [3]), whether this also applies to GUR needs yet to be determined.

Recommendations

Based on our experience we can recommend the following procedures:

1. Make sure you reduce the participants pool to your target audience
2. Combine quantitative measures with qualitative data for a more detailed understanding
3. Include one or multiple checks to reduce careless responses such as bogus items (e.g., "Respond with 'strongly agree' for this item") or self-report measures of data quality (e.g., "In your honest opinion, should we use your data?") (refer to [6] for an extensive review of careless response detection methods).
4. Pay only a small amount for the completion of the survey and give participants who responded carefully a bonus
5. Examine data quality by combining objective measures such as outliers in response time with reported technical difficulties

Conclusion

Crowdsourced playtesting is an ideal tool for indie, nonprofit and academic GUR. It provides access to a large and diverse user base that can be used to receive cost efficient and timely feedback from players.

References

[1] Judd Antin and Aaron Shaw. 2012. Social Desirability Bias and Self-reports of Motivation: A Study of Amazon Mechanical Turk in the US and India. In *CHI '12*. ACM, 2925–2934.

[2] Florian Brühlmann and Gian-Marco Schmid. 2015. How to Measure the Game Experience? Analysis of the Factor Structure of Two Questionnaires. In *CHI '15 EA*. ACM, 1181–1186.

[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.

[4] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *CHI '08*. ACM, 453–456.

[5] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[6] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods* 17, 3 (2012), 437.

About the authors

Florian Brühlmann and Gian-Marco Schmid are PhD students and Elisa Mekler is a postdoctoral researcher at the University of Basel HCI research group. Their research interests include motivational and emotional processes of the player experience and their measurement.