"Are Crowdsourcing Platforms Reliable for Video Game-related Research?" A Case Study on Amazon Mechanical Turk

Linus Eisele
linus.eisele@uni.li
Liechtenstein Business School
University of Liechtenstein
Liechtenstein, Vaduz

Giovanni Apruzzese
giovanni.apruzzese@uni.li
Liechtenstein Business School
University of Liechtenstein
Vaduz, Liechtenstein

Abstract

Video games are becoming increasingly popular in research, and abundant prior work has investigated this domain by means of *user studies*. However, carrying out user studies whose population encompasses a large and diverse set of participants is challenging. Crowdsourcing platforms, such as Amazon Mechanical Turk (AMT), represent a cost-effective solution to address this problem. Yet, prior efforts scrutinizing the data-quality (unrelated to gaming) collected via AMT raises a concern: is AMT reliable for game studies?

In this paper, we are the first to tackle this question. We carry out three user studies (n=302) through which we evaluate the overall validity of the responses—pertaining to 14 popular video games—we received via AMT. We adopt strict verification mechanisms, which are trivial to "bypass" by real gamers, but costly for non-gamers. We found that the percentage of valid responses ranges from 5% (for WoW) to 28% (for PUBG). We hence advocate future research to carefully scrutinize the validity of responses collected via AMT.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → **Media arts**.

Keywords

mechanical turk, game studies, league of legends, world of warcraft, crowdsourcing, mturk

ACM Reference Format:
Linus Eisele and Giovanni Apruzzese. 2024. "Are Crowdsourcing Platforms Reliable for Video Game-related Research?" A Case Study on Amazon Mechanical Turk. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY Companion '24), October 14–17, 2024, Tampere, Finland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3665463.3678817

1 Introduction

The video game industry is in constant expansion, both in terms of revenue (with an expected value of over $500 billion in 2027) and pervasiveness: as of 2022, over 40% of the global population plays video games [29]. Nowadays, "video gamers" are scattered in every corner of the world, and thanks to the advances of information and communication technologies, players (of all ages and groups) can now benefit from a variety of games—which they can consume anywhere and at any time [17, 29, 44, 71].

Alongside the growth of video games as a media, also video games *research* has substantially evolved over the years. In particular, abundant papers have investigated the connections between our "real and virtual" lives through video games—such as gambling [15, 45], or our behaviour during social interactions in online role playing games [47, 68]; some even scrutinized how games may shift our own values during our everyday lives [41]. Simply put, it is now acknowledged that video games are not just a form of entertainment, and there are many aspects of our nature that can be discovered (or, potentially, enhanced [55]) by carrying out human-subject research centered around video games.

In this context, we observe that this line of research (typically known as "game studies") has also undergone some changes from a methodological perspective: various prior work (e.g., the above-mentioned ones [15, 41, 45, 47, 68], as well as others [28, 64]) are now relying on *crowdsourcing* methods to recruit the participants through which to investigate a certain game-related hypothesis. Such a practice is well-justified: gathering a representative (i.e., large and diverse) sample that can approximate the world's population of gamers is daunting, and crowdsourcing platforms, such Amazon Mechanical Turk (AMT), are a convenient choice to circumvent this problem. Indeed, by investing a certain sum of money (to be paid after the completion of a given user survey), one can easily collect a pre-defined set of responses—provided remotely and stemming from individuals located across the globe. In this way, user studies counting >1000s of participants can be carried out [13].

Unfortunately, crowdsourcing platforms are not exempt from issues, and various prior work (e.g., [61]) have questioned their overall utility for research; for instance, the very recent work by Christoforou et al. [18] revealed that "generative AI" is widely used by the population of well-known crowdsourcing platforms. However, we found no evidence of any work that specifically investigated the reliability of crowdsourcing solutions in the video-game domain. This is a problem: if researchers are going to recruit participants via crowdsourcing for game-related purposes, it is paramount to determine if such solutions can yield usable results—and, if this is not the case, devise appropriate mitigations that ensure a smooth development of future research on video games. To the best of our knowledge, we are the first to undertake this challenge.

Contributions. In this Work-in-Progress (WiP) paper, we seek to promote more reliable research on game studies. To this

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI PLAY Companion '24, October 14–17, 2024, Tampere, Finland
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0692-9/24/10
https://doi.org/10.1145/3665463.3678817

end, after providing the necessary context to appreciate our efforts and define the scope of our work (in Section §2), we:

- design and realise three surveys (one focused on League of Legends, another on World of Warcraft, and a third "generic" one focused on 14 games) which we distribute on AMT and which entail strict validatory mechanisms (§3);
- after collecting 302 responses, we analyse their validity (§4). Our quantitative results underscore that only a tiny fraction (less than 13% overall) of our responses are valid—much less than that reported by some prior related work.
- We then discuss and outline the limitations of our study (§5), and derive recommendations for future work (§6). Importantly, *we do not (nor aim to) invalidate* prior work, and we do not claim that AMT cannot be used for research on game studies.

To ensure scientific replicability of our research, we provide some of our resources in a dedicated repository [3].

2 Background and Related Work

To set up the stage for our contribution, we position our paper within existing literature (§2.1), outline the fundamental concepts of crowdsourcing platforms (§2.2), and motivate the necessity of our research (§2.3).

2.1 Game Studies and Human-subject Researches [Focus of the Paper]

Video games have garnered abundant attention in research, and prior work has looked at the (video) gaming domain from a variety of perspectives (e.g., [20, 23, 30, 55, 56, 58, 59, 59]). In this paper, we focus on *game studies*, i.e., that branch of research which seeks to analyse the interplay between a game and the surrounding environment—typically by means of user studies (e.g., interviews or surveys) entailing real people. Topics of study include: psychological [38], personality [14] or social aspects [67] (including, e.g., purchasing behaviours [40]); personification [43] or gamification [32]; improvement of in-game performance [52] or quality of experience [31] of video-gamers; as well as security and privacy [66] of players. Potentially, such studies can entail gaming communities gravitating around third-party platforms (e.g., Twitch [46]).

The common characteristic of this broad category of research works is the necessity of finding (or recruiting) a suitable number of *human subjects* that can participate in the corresponding user study: Intuitively, studies involving populations that are both large and diverse allow deriving more generalizable (or well-founded) conclusions. However, meeting such requirements is not trivial from an organizational perspective: for instance, people may be unwilling to participate "for free," and even though it is possible to provide some incentive (e.g., a monetary reward or gift [35]), there is still the problem of advertising/distributing the task so that a sufficient number of individuals is reached.

2.2 Crowdsourcing Platforms [Technical Background]

As a necessary digression, let us outline the landscape of *crowdsourcing* platforms, whose growth has substantially increased in the last decade, and are now considered as being an established "data collection method" in research [25, 33, 69]. A prominent example of crowdsourcing platforms[1] is Amazon Mechanical Turk (AMT), whose rollout began in 2005 [1] and which counts hundreds of thousands of users [22, 60]. On AMT, *workers* perform on-demand tasks, referred to as *HIT* (short for Human-Intelligence Task), receiving monetary compensation paid by the *requester* for the successful completion of such tasks [51]. In other words, the "wage" of a given worker depends on the number of HITs they complete: the faster a worker can complete any given HIT, the higher their overall revenue from AMT. Hence, workers are incentivized to be fast.

Prior work [34, 51] estimated an hourly wage for AMT workers to be in-between 2–5$, but recent efforts quantified it as being potentially much higher [61], up to 25$ per hour (before taxes) for "quick" workers. Indeed, the majority of HITs are repetitive microtasks that may take seconds or few minutes to complete, such that workers can quickly submit their responses and move on to participating to a new HIT to increase their earnings. To ensure that workers are only paid if they do a good job (i.e., the HIT is completed truthfully and fairly), requesters are given the possibility of approving (or rejecting) the HIT received for any given *batch* (i.e., a set of HITs pertaining to the same purpose): workers whose HIT is accepted/rejected will be paid/not-paid their due compensation—which is specified upon the creation of the batch. On this note, AMT offers a variety of settings to enhance the quality of the responses collected for any given batch. For instance, requesters can specify a minimum number of HITs that workers must have completed successfully in order to be eligible for their batch, as well as a minimum approval rate of a worker's overall completed HITs.

Crowdsourcing platforms, such as AMT (but also, e.g., Prolific or Qualtrics [24]), have been historically known to be particularly suited for tasks related to the development (or assessment) of techniques within the artificial intelligence domain [37]: for instance, the popular ImageNet dataset was labeled by AMT workers [49], and some prior works also used AMT to validate their proposal (e.g., [5, 62]). However, these platforms have been also used for behavioral and human-subject research (e.g., [4, 61]), including papers related to (video) game studies.

2.3 Amazon Mechanical Turk and (Video) Game Studies [Motivation]

As we stated, carrying out game-related user studies is challenging, and crowdsourcing represents an enticing and convenient solution to such a challenge. Indeed, platforms such as AMT allow streamlining the data collection procedure, since a researcher (as a "requester") only needs to *(i)* devise a certain form/questionnaire to investigate a given aspect; *(ii)* set up the batch by configuring a few parameters, such as number of responses and reward per HIT; and *(iii)* publish the batch—and then, after reviewing the results, potentially repeat this process anew to collect new responses. In light of such simplicity, it is not surprising that many papers have leveraged various crowdsourcing tools to carry out their user studies. To provide some examples, in the last five years (since 2019), AMT has been leveraged by: Brooks and Clark [15] and Wang et al. [68] in 2019; Wohn et al. [70] and Jang and Byon [36] in 2020; Beres et al. [11] and Kelly et al. [39] in 2021; Kowert et al [41] and Deaner et

[1] We note that "crowdsourcing" stems from "work outsourcing" [2], i.e., participants are paid—which is different from disseminating surveys on social networks: such a complementary way of carrying out user studies (e.g., [12, 63, 66]) is outside our scope.

al. [21] in 2022; Larche et al [45] and Li Anthony et al. [47] in 2023. Moreover, there are also many game-related works that have relied on AMT in the last decade prior to 2019, such as [13, 28, 38, 42, 64].

Problem Statement (and Research Gap). Despite the many works that have relied on AMT for research purposes, we observe that some prior endeavours criticized the overall reliability of AMT for research purposes. For instance, a 2021 paper from Saravanos et al. [61] found that over 33% of responses collected via AMT are unreliable. In 2022, Agley et al. [7] highlighted the importance of implementing strict quality-control mechanisms for AMT. More recently, Marshall et al. [50] investigated the reliability of AMT over a 10-years timespan, and found that unusable data increased from as little as 2% in 2013 to over 90% in 2022. Yet, all these (and others, such as [6, 18, 48, 65]) "critical analyses" *did not focus specifically on (video) game studies*. Hence, motivated by the many papers that rely on crowdsourcing for the video game domain, we seek to scrutinize the reliability (in 2024) of AMT for game-related research. We specifically ask ourselves: "is AMT a reliable platform for user studies wherein participants are expected to be video-gamers?" We stress that this is a WiP paper: as we will discuss (§5), providing a clear answer to such a question is intrinsically difficult.

3 Research Methodology

To investigate our research question, we use AMT to carry out three user studies focused on popular and recent video games, and then assess the ratio of "valid" responses we collect. Hence, in this section, we first describe the rationale behind our chosen video games (§3.1), explain how we designed our questionnaires (§3.2), and finally present how we configured the AMT platform (§3.3), which we used as "requesters".

3.1 Video Game Selection

To provide results that are more representative of the current video-gaming landscape, we carried out three user studies: one focused on League of Legends (LoL), one focused on World of Warcraft (WoW), and one focused on 14 popular games with a multi-player component. Let us justify our choices.

- League of Legends is one of the most popular MOBA games, having over 140M active monthly players, with an average of 900K concurrent players. We chose LoL as a representative candidate for "competitive" games. Besides, LoL has been considered by prior work carrying out user studies on AMT (e.g., [68]).
- World of Warcraft has been the most popular MMORPG since 2004, and its population is very diverse. As of May 2024, WoW has over 30M active monthly players and nearly 250K concurrent players. We chose WoW as a representative candidate for "cooperative" games. Besides, WoW has also been widely considered in prior work carrying out user studies on AMT (e.g., [21]).
- The last "generic" user study spans over 14 games: Counter Strike 2 (CS2), Rocket League, Fortnite, PlayerUnknown's Battlegrounds (PUBG), Apex Legends, GTA Online, Battlefield 2042 (BF), Overwatch 2 (OW2), Valorant, Call of Duty: Warzone (CoD:W), Rainbow Six Siege (RSS), Destiny 2; as well as WoW and LoL, Altogether, these titles represent a mix of competitive or cooperative video games, which are regularly played by millions of players worldwide.

Statistics about the playercount are taken from: activeplayer.io, playercounter.com, and steamcharts.com (on May 2024).

We carry out three user studies for a simple reason: we do not know *what games are played by AMT's "workers"*. Indeed, our research is exploratory in nature: to the best of our knowledge, there is no prior work that assessed if AMT can be considered a suitable environment for any of the user studies we seek to carry out (after all, we seek to shed some light on the validity of AMT). Hence, in the last "generic" user study, we will ask participants to choose the game they play the most (among our list of 14 games), thereby allowing us to discern if there is a specific game for which AMT can be considered as a better reservoir of valid candidates for game-related user studies.

3.2 Survey Design and Implementation

At a **high-level**, each user study follows a similar workflow: participants (i.e., AMT workers) are brought to a questionnaire (hosted on Google Forms)[2] involving a mix of simple questions, ideally resembling a "typical" survey. Specifically:

- *Introduction:* we summarize our survey, provide our institutional contacts for inquiries, and ask for the WorkerID. We also inform the participant that we will not share their data (which we treat with confidentiality) with anyone.
- *Demographics:* we ask closed questions about age, country, gender, employment status (as done in, e.g., [38, 40]).
- *Personality:* we ask 10 closed questions for the "Big-Five Personality Traits" (common in related research [14, 66]).
- *Game-related:* we ask 3–8 questions (depending on the game) about the game of choice. The questions are simple and any "regular" player of any of the games we considered should be able to answer these right away.
- *End:* we provide a "survey code" that the participant must input on AMT to conclude the HIT.

Importantly, for the third "generic" user study, we ask a preliminary question inquiring "which game [among the listed 14 games] do you play the most?": depending on the answer, the participant will be brought to a specific section of the questionnaire, focused on the chosen game. All these questionnaires are in our repository [3].

The crucial part of our questionnaire, however, is the one devoted to the game-related questions. Indeed, it is here that **we assess the trustworthiness of the responses received**. We do so by means of three types of validatory mechanisms (inspired by [68]):

- *Publicly-available Gamertag.* We ask for the participant's gamertag, which should be provided as a link (i.e., a string) to a publicly accessible website (e.g., op.gg) through which we can validate whether the provided input is genuine. In phrasing the question, we also provide an exemplary link (pointing to, e.g., the profile of one of the authors, or of a professional player) to facilitate the understanding of the format of the answer we seek to receive.
- *Player-related.* We ask for information about the participant's in-game activity, which we can verify by checking the provided link. For instance, we can ask "what is your favourite weapon?" or "what is your most played hero?" or "what is your favourite role?" or "what is your goal/shot ratio?".

[2] In the questionnaire we do not collect information that precisely reveals the real identity of an individual (e.g., name or physical address, and not even the email).

- *Game knowledge.* For the generic survey, we ask trivial questions about the user-chosen game. For instance, for Valorant, we ask "what is the name of the bomb-type device in Valorant?" and the possible answers are "Buster", "Sprint", "Blitz", "Spike".

A response is genuine *if and only if* all the following are true: *(i)* the participant provides a valid gamertag, i.e., the link points to the profile of a player that is publicly visible; *(ii)* the answer to the player-related information matches the data shown in the provided link; *(iii)* the answers to the game-knowledge questions are correct; *(iv)* the survey code inserted on AMT matches the one we specified at the end of the questionnaire. If any of these is wrong, then the response is unusable, and we will reject the HIT on AMT; otherwise, the response is usable and we will approve the HIT. Validation was done via custom-made scripts (e.g., checking the URL) as well as manually (e.g., for browsing the profile).

Ethical and Transparency Statement. Our institution, despite being aware of our research, does not have any formal IRB process. However, we follow ethical principles for our study [10]. First, our survey does not collect any sensitive information [19, 53] about our participants. Second, the information we collect (which is essentially the same collected by [68]) is not explicitly prohibited by AMT [8, 9]. We acknowledge that the WorkerID and the (public) gamer profile can be used to identify a person if cross-correlated with other information [54] (not collected via our survey), which is why such data must be treated with caution. Hence, to comply with the GDPR [27] and also with AMT's policies [8]: *(i)* we do not try to infer more personal details about our participants, *(ii)* we provide our contacts for inquiries, *(iii)* no information is collected until the participant presses the "submit" button, and *(iv)* we deleted all received data. Third, to avoid deception, we are transparent in informing the participants that *the survey will require the participant to provide their gamertag "as an URL to their gamer profile"*, and that *failure to do so will lead to rejection of the HIT*. Specifically, we write this in *(i)* the description of the batch; *(ii)* the AMT page of the batch; *(iii)* at the beginning of the Google Form; *(iv)* in the question where we ask for the gamertag. Hence, our participants are well-aware that the task requires providing this (publicly available) data. Fourth, no harm is done to our participants; our survey is short (≈5m) and even in cases of mistakes (or second thoughts) a participant would not have wasted a considerable amount of time. Finally, to protect the privacy of our participants, we will not reveal any specific response or result (which are outside our scope) beyond those pertaining to the validity of the answers we have received. (The "Demographics" and "Personality" questions in our questionnaire served to ensure broad applicability of our findings: most surveys ask for these.)

3.3 AMT Configuration and Batch Setup

Once we crafted our questionnaires, we set up the corresponding batches on AMT. To ensure a consistent testbed, we set the same configuration parameters for all user studies. Specifically, we chose two filters: "HIT Approval Rate>98%" and "Number of approved HITs>5,000"; these criteria are chosen based on established practices [26, 61]. The reward per HIT was set to 1.5$: we carried out pilot studies and estimated 5m to complete each survey, meaning that our compensation was above the "acknowledged" average hourly rate of similar platforms [57, 61] (besides, fast workers could complete this in half the time, potentially making our task worth up to 30$ per hour). We set a duration for the task of 15 minutes to allow workers to take their time. We set the number of responses received (per batch) to 100.

We provide a screenshot of the "preview" provided by AMT to the LoL user study in Fig. 1, showing the configuration parameters mentioned above, as well as the textual description of the task (shown to workers). We provide the pages showing the exact configurations for each of the user studies in our repository [3].

4 Key Findings and Results

We carried out our user studies in the first week of May, 2024. On average, the batches have been completed after ~3 hours since their publication. We first present the quantitative results (§4.1), which we then compare to those of prior work (§4.2), and finally provide a qualitative analysis underscoring relevant findings (§4.3).

4.1 Quantitative Results

We report the detailed results of our three user studies in Fig 2. It stands out that the number of "invalid" responses significantly exceeds that of "valid" responses (any statistical test would confirm this hypothesis with $p \sim 0$). In particular, for WoW, only 5% of our responses are valid, whereas the percentage is higher for LoL (16%) and for the "generic" survey (17%), for which we provide per-game breakdown[3] of the responses in Table 1. We see that, for PUBG, 12 responses out of 42 (i.e., 28%) are valid, which is significantly higher than any other game (if we exclude the results obtained for LoL in the generic study, for which we did not have enough samples to derive statistically sound conclusions). Also notable is that RocketLeague ranks second among the most "popular" games played by our participants.

Takeaway: On a per-game basis, at most 28% responses (PUBG) are valid, with drops to as low as 5% (WoW). The average is 13%.

4.2 Comparison with Prior Work

Let us compare our findings with those achieved by prior work that used AMT for their user studies—but whose surveys *were not designed to test our hypothesis*.

First, we begin by mentioning the findings of the 2019 paper by Wang et al. [68] (focusing on LoL), which also relied on the gamertag for verification but (as far as we are aware) did not overly-emphasize in the HIT that its validation depended on the provided gamertag: the results of [68] showed that nearly 30% of responses from AMT were usable—almost 2x ours (for LoL). Then, we mention [36], which considered various esport games and set up their AMT batch by specifying >500 approved HITs with >97% HIT rate, but only relied on "attention checks" and on "completion time" as validation mechanisms: the findings of [36] revealed that 87% of the 1,129 received responses were valid. Similarly, Brooks and Clark [15] (who set up AMT by specifying >1,000 approved HITs with > 98% HIT rate) found that 85% of the 1,000 responses were valid. Finally, Larche et al. [45] focus on Overwatch and set up their AMT batch by specifying >1,000 approved HITs with >96% HIT rate, and found

[3] Due to how Google Form works, some workers completed the form *after the batch was closed on AMT*, leading to us receiving 103 answers instead of 100.

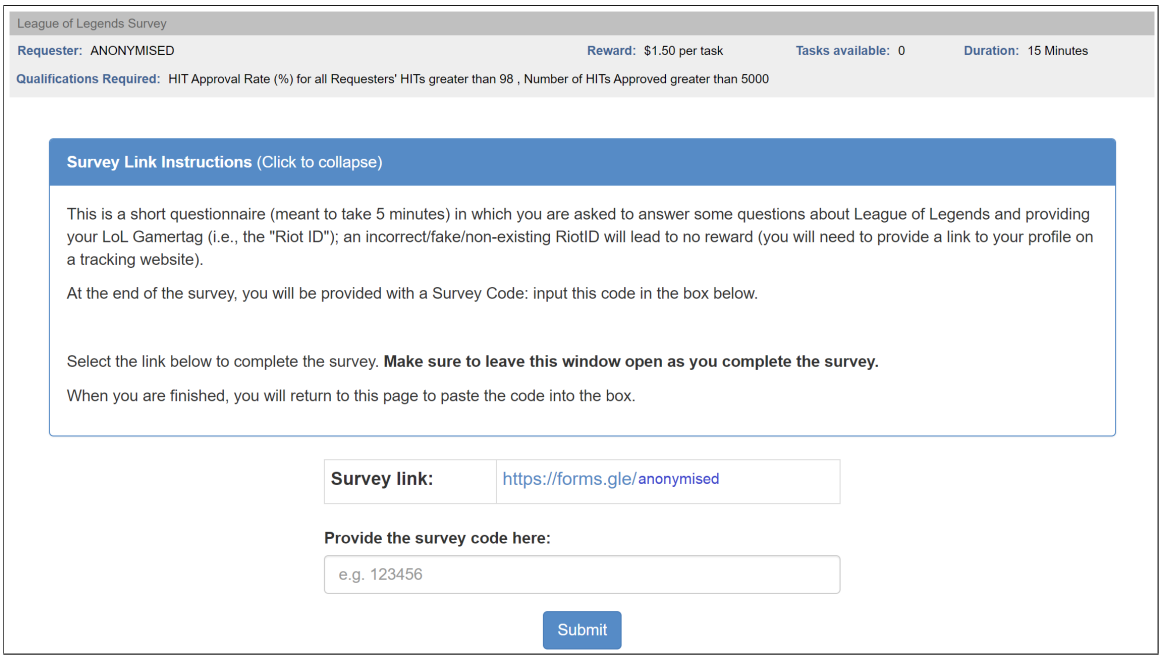

Fig. 1: AMT Configuration for the LoL User Study – We set restrictive parameters (according to [26, 61]) to ensure high-quality answers. The description clearly specifies that we request the participants to input their gamertag as a link to a tracking website, or no compensation will be paid.

that 58% of the 438 received responses were valid: accordingly, "All potential participants must have played Overwatch at least once in the past 4 weeks [...] Eligibility was established via a [...] prequalifcation questionnaire" (albeit we are unsure of the exact validation mechanism employed in such a questionnaire).

In summary, the percentage of valid responses in the above-mentioned prior related work – which relied on different, and less strict, validatory mechanisms – was superior to ours.

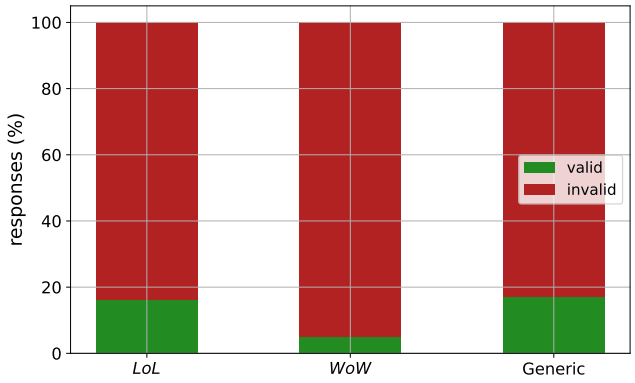

Fig. 2: Main results. Ratio of valid/invalid responses for each survey.

4.3 Qualitative Analysis

We provide some insights on the responses we collected. First, **the major reason for invalid responses was a fake (or incorrect) gamertag**. Many respondents provided the same URL we put in the example; others provided a similar URL by changing just one or two characters. Many provided a (fake) gamertag of an account that "made sense" (e.g., we had a very high number of "EarlGreyTeemo" for League of Legends). In some intriguing cases, the provided gamertag did not match the URL of their profile. In a reduced number of cases, the gamertag and URL were correct, but the answers to the "player-related information" were incorrect: for instance, one answer provided the URL and gamertag of a high-ranked player, but the favourite weapon was incorrect; of course, during our validation, we did account for potential "changes" (e.g., a weapon may have been become the "favourite" only recently). In some cases we received (correct) responses pertaining to high-ranked players, which we considered as valid since we have no way to disprove their authenticity (aside from direct contact to the worker or via invasive analyses—which we could not do for ethical reasons.).

Video Game	✓	✗
PUBG	12	30
RocketLeague	1	11
CoD:W	0	9
RSS	0	7
GTA:Online	0	7
WoW	0	5
Destiny2	1	4
LoL	2	2
CS2	1	3
BF2042	0	2
OW2	0	2
ApexLegends	0	2
Valorant	0	2
Fortnite	0	0

Table 1: "Generic" survey breakdown. Ratio of valid/invalid responses for each survey. We report the number of valid (✓) and invalid (✗) responses for the specific games selected in the "generic" user study.

Due to the above, *we conjecture that the majority of these tasks have been done through automated* (e.g., Generative AI [18]) means—as evidenced by many responses being similar to each other (e.g., the abovementioned "EarlGreyTeemo") or the mismatched link/gamertag (which any real player of these games would know how to bypass).

We also mention that we have received follow-up messages from two workers of rejected HITs. These messages asked us to reconsider our decision to reject their HIT. We were willing to do so, and politely responded that their HIT was rejected due to an incorrect/fake gamertag. We did not receive any response confirming the "legitimacy" of the input.

5 Discussion and Lessons Learned

We provide additional results from some prior experiments we had carried out (§5.1) Then, we discuss the pros-and-cons of the methods (including our own) used to validate responses in user studies (§5.2) and outline the limitations of our research (§5.3).

5.1 Backstory (and preliminary investigation: a negative experiment)

To further substantiate our findings, we explain *why we carried out the research described in this paper*. Indeed, aside from the "objective" motivation presented in §2.3, there is a "subjective" reason that inspired us to investigate the reliability of AMT for game studies.

In April 2024 we sought to use AMT to carry out a game-related user study. We followed a similar procedure as described in §3. We created a questionnaire entirely focused on LoL (since that was the game we were investigating at the time), having a similar structure as the one discussed in §3.2. The only difference was that we did not excessively emphasize the importance of providing a gamertag: we only wrote about it on the HIT description on AMT, and at the beginning of our questionnaire. We also provided slightly different example strings for the URL. As for the configuration on AMT, we created four batches: the first had a single restriction of "HIT approval rate >92%", and we set a batch size of 150; for the second, we added the constraint "number of HITs approved >50", and set the batch size to 50; for the third, we set the same constraints as in §3.3, but the batch size was 25; the fourth was a more restrictive one for which we enabled the "Master Worker" filter.

For the first batch, all 150 HITs were rejected (here, the number of "EarlGreyTeemo" was very high); the same applies for the second batch: all 50 HITs were rejected. For the third batch, we approved 2 HITs and rejected 23. For the fourth batch, we received only 6 HITs (the batch never reached the target number of 25) which were all invalid. Hence, out of 231 responses (for LoL), we only had 2 useful ones. Our "negative experiment" surprised us so much that we set aside our initial goal and decided to focus on this research, in which we further clarified (to the extent of potentially overemphasizing it) that our surveys require to provide the gamertag and the URL to the user profile for validation purposes to test our hypothesis.

5.2 Tradeoffs: a matter of Trust [Reflection]

When carrying out user studies, *researchers implicitly assume an element of trust*, in that participants fulfill the tasks truthfully and honestly. Typical validation mechanisms are often meant as "attention checks" (e.g., [13]) whose purpose is ensuring that participants do not skip certain parts of a survey by providing superficial answers. These mechanisms are crucial to ensure reliable findings of any user study—but they may not be enough for crowdsourcing.

Indeed, our findings (which echo those of works in other domains [6, 18, 48, 50, 61, 65]) revealed that most of the responses we solicited from AMT workers who claim to be "gamers" are unlikely to be generated by "real gamers." To provide such findings, we resorted to a strict verification mechanism wherein we require participants to submit their gamertag (as also done by [68]). Such a mechanism is, however, a double-edged sword: on the one hand, it allows for unambiguous verification of "invalid" responses (i.e., non-existing gamertags); on the other hand, it may not be adopted universally (e.g., some games do not have any "gamertag" that can be used for such verifications; moreover, collecting gamertags requires delicate ethical considerations) and it can still be circumvented (e.g., by providing a valid gamertag that does not belong to the respondent).[4] We hence endorse to exercise caution when considering the application of our methods.

Prior work has adopted various methods to address the issue underscored in our paper. For instance, a workaround are "screening" mechanisms to identify qualifying participants *before* carrying out the actual study.[5] This is done, e.g., in [15, 21, 36, 45]. However, such screening tasks can also be completed untruthfully: for instance, the screening in [36] is done via "a screening question to determine whether participants had experience playing esports games" whereas [15] described the screening as "easblish[ing] eligibility included prior video game play and familiarity with loot boxes".[6] Unfortunately, it is challenging to ensure the trustworthiness of such responses without affecting the users' privacy. Intriguingly, Kelly at al. [39] ask participants to provide "a photo of their device", which may still leak private details.

In summary, every approach to ascertain the trustworthiness of the responses collected for game-related user studies presents **privacy/utility tradeoffs**. Asking many validatory questions and/or requiring certain information to be provided may increase the validity of the findings, but it may discourage some people from participating (and may also pose ethical concerns which must be addressed). In contrast, determining the eligibility of participants via self-assessments is convenient, but it may not filter out responses that should be excluded to avoid noisy data.

5.3 Limitations and Scope

Our research has a number of limitations. First, even though we have carried out three user studies entailing hundreds of people, *our results cannot prove that the entirety of AMT's workforce is unreliable* (or unsuitable) for game studies' research.

[4] We note, however, that such "circumventions" are not trivial to apply by non-gamers, since their application would increase the time to complete the HIT substantially (given that finding a "valid gamertag" is not straightforward without knowledge of the considered game). This may discourage their adoption by AMT workers, who may deem the HIT to be not as rewarding as initially expected).

[5] It is also possible to determine eligibility by disseminating a given survey to some specific communities (e.g., social media) and then ask members of such community to participate to the user study and that compensation may be given upon completion of the survey. However, such a mechanism would defeat the purpose of using crowdsourcing services such as AMT, whose main utility is to provide a stand-alone platform that facilitates recruitment, completion and payment.

[6] We found no details on whether such screening entailed verification mechanisms similar to those used in our paper (e.g., asking the gamertag and checking the public profile), which is why we assume that such screening relied on self-reported assessments.

For instance, our user studies have been carried out during a short timeframe, and it is possible that if we had submitted our batches during different times/days, we would have collected more valid responses. Moreover, it may be that our transparency backfired: some workers may not have participated in the survey due to the explicit mention of the gamertag (which they did not want to provide) in the HIT's description. Another shortcoming is that we have only considered 14 games: despite being popular titles, there may be other games for which AMT may be more reliable.

Finally, we have only considered AMT: other crowdsourcing platforms (e.g., Qualtrics or Prolific—the latter not allowing Gamertags by default) have a different userbase, and hence we cannot make any claim on whether our conclusions apply to the entire spectrum of crowdsourcing solutions. Investigating these complementary settings (which have been used in related research [16]) and assessing whether they are more/less suitable than AMT is a valuable path for future work.

6 Conclusion and Recommendations

Through three user studies focusing on 14 popular video games, we have discovered that AMT may not be "blindly" used to carry out valuable research for game studies. Such a finding should induce our community to reflect upon the usage of existing crowdsourcing platforms. We take the first step, and derive three takeaways.

First and foremost, a disclaimer. We carried out our research in May 2024. As such, even though our findings underscored that AMT is "not-very-reliable" (for game studies), we cannot (nor want to) make any claim with regards to the validity of game-related user studies carried out by prior research. Put differently: our findings *do not invalidate the results of prior work* that carried out user studies on AMT to investigate a given phenomenon.

Second, a lessons learned is that *validation is crucial*. We designed our questionnaires by integrating various validation mechanisms to ensure accurate verification of the responses. We posit that researchers should come up with ways (not necessarily entailing the gamertag) that could not be bypassed via automated mechanisms (e.g., ChatGPT) since our findings suggest that some workers may be relying on similar tools for their HITs.[7] To this end, we recommend leveraging domain expertise in the gaming genre to insert questions that an AI could not answer properly, or which would require a substantial amount of time to answer if the participant is not truly a player of the considered game.

Third, and last: a constructive solution to the problem we identified is the *creation of a crowdsourcing platform specific of gamers*. Given the growth of this research field, we believe that the gaming community would greatly benefit from the introduction of a similar solution. For instance, gamers can register (as "workers") by providing their in-game details, and then can easily participate in various "HITs" (submitted by researchers, i.e., "requesters") revolving around some game-related aspect. In this context, the platform should provide the verification mechanisms that ascertain that any given worker is "a player of a certain game" (potentially by providing additional details, e.g., competence level), thereby allowing requesters to selectively carry out user studies targeting a specific population of gamers. This may require a higher payment, but we believe that such additional cost is worth the price, since it would lead to more correct (and, hence, valuable) research. We hope our findings can kickstart future endeavours focusing on developing such a solution, and we will commit to this effort.

Acknowledgement. We thank: the organizers (chairs, meta-reviewer, reviewers) of the WiP track of CHI PLAY 2024 for their constructive comments which improved this paper immensely; and the Hilti group for funding this research.

[7] An orthogonal, but intriguing, research question that naturally occurs from our findings is: "how it is possible that over 80% of our participants have such a high HIT rate despite failing our validation?" According to [18], AMT workers may use GenAI, which appears to be proficient at solving HITs—but [18] did not focus on game studies.

References

[1] 2024. Amazon Mechanical Turk (Wikipedia). https://web.archive.org/web/20240517120039/https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk.
[2] 2024. Crowdsourcing (Wikipedia). https://web.archive.org/web/20240715201145/https://en.wikipedia.org/wiki/Crowdsourcing.
[3] 2024. Repository of this paper. https://github.com/hihey54/chiplay24.
[4] Tahir Abbas and Ujwal Gadiraju. 2022. Goal-setting behavior of workers on crowdsourcing platforms: An exploratory study on mturk and prolific. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 2–13.
[5] Hadi Abdullah, Muhammad Sajidur Rahman, Washington Garcia, Kevin Warren, Anurag Swarnim Yadav, Tom Shrimpton, and Patrick Traynor. 2021. Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 712–729.
[6] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Fabian Flöck, and Jens Lehmann. 2018. Detecting linked data quality issues via crowdsourcing: A dbpedia study. *Semantic web* 9, 3 (2018), 303–335.
[7] Jon Agley, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. 2022. Considerations for conducting alcohol research with the USAUDIT on Mechanical Turk (mTurk). *Journal of studies on alcohol and drugs* 83, 1 (2022), 159–161.
[8] Amazon. 2018. *MTurk's Acceptable Use Policy*. https://web.archive.org/web/20240712184710/https://www.mturk.com/acceptable-use-policy
[9] Amazon. 2020. *Participation Agreement*. https://web.archive.org/web/20240712184709/https://www.mturk.com/participation-agreement
[10] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The Menlo report. *IEEE Security & Privacy* (2012).
[11] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proc. CHI*. 1–15.
[12] Kelly Bergstrom and Nathaniel Poor. 2021. Reddit gaming communities during times of transition. *Social Media+ Society* 7, 2 (2021).
[13] Max V. Birk, Maximilian A. Friehs, and Regan L. Mandryk. 2017. Age-Based Preferences and Player Experience: A Crowdsourced Cross-sectional Study. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Association for Computing Machinery, 157–170.
[14] Beate Braun, Juliane M Stopfer, Kai W Müller, Manfred E Beutel, and Boris Egloff. 2016. Personality and video gaming: Comparing regular gamers, non-gamers, and gaming addicts and differentiating between game genres. *Computers in Human Behavior* (2016).
[15] Gabriel A. Brooks and Luke Clark. 2019. Associations between loot box use, problematic gaming and gambling, and gambling-related cognitions. *Addictive Behaviors* 96 (2019), 26–34. https://doi.org/10.1016/j.addbeh.2019.04.009
[16] Xiaowei Cai, Javier Cebollada, and Mónica Cortiñas. 2022. Self-report measure of dispositional flow experience in the video game context: Conceptualisation and scale development. *International Journal of Human-Computer Studies* 159 (2022), 102746.
[17] Isaac Cheah, Anwar Sadat Shimul, and Ian Phau. 2022. Motivations of playing digital games: A review and research agenda. *Psychology & Marketing* (2022).
[18] Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. 2024. Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 2097–2103.
[19] European Commission. 2024. Sensitive Data. https://web.archive.org/web/20240204163656/https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en.
[20] Gillian Dale and C Shawn Green. 2017. The changing face of video games and video gamers: Future directions in the scientific study of video game play and cognitive performance. *Journal of Cognitive Enhancement* 1 (2017), 280–294.
[21] Robert O Deaner, Lucretia C Dunlap, and April Bleske-Rechek. 2022. Sex Differences in Competitiveness in Massively Multiplayer Online Role-Playing Games (MMORPGs). *Evolutionary Psychology* 20, 2 (2022), 14747049221109388.

[22] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 135–143.

[23] Mary Jo Dondlinger. 2007. Educational video game design: A review of the literature. *Journal of applied educational technology* 4, 1 (2007), 21–31.

[24] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one* 18, 3 (2023), e0279720.

[25] Adam Enfroy. 2024. *7 Best Crowdsourcing Platforms of 2024 (Ultimate Guide)*. https://www.adamenfroy.com/crowdsourcing-platform

[26] Finding Five. 2020. Tutorial: Mechanical Turk Best Practices. https://web.archive.org/web/20240225025247/https://news.findingfive.com/2020/05/17/mechanical-turk-best-practices/. Accessed: May, 2024.

[27] GDPR.eu. 2019. *What is considered personal data under the EU GDPR?* https://web.archive.org/web/20240513154019/https://gdpr.eu/eu-gdpr-personal-data/

[28] Erica A Giammarco, Travis J Schneider, Julie J Carswell, and William S Knipe. 2015. Video game preferences and their relation to career interests. *Personality and Individual Differences* 73 (2015), 98–104.

[29] Edward Goh, Omar Al-Tabbaa, and Zaheer Khan. 2023. Unravelling the complexity of the Video Game Industry: An integrative framework and future research directions. *Telematics and Informatics Reports* (2023), 100100.

[30] Isabela Granic, Adam Lobel, and Rutger CME Engels. 2014. The benefits of playing video games. *American psychologist* 69, 1 (2014), 66.

[31] David Halbhuber, Niels Henze, and Valentin Schwind. 2021. Increasing player performance and game experience in high latency systems. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–20.

[32] Stuart Hallifax, Maximilian Altmeyer, Kristina Kölln, Maria Rauschenberger, and Lennart E Nacke. 2023. From Points to Progression: A Scoping Review of Game Elements in Gamification Research with a Content Analysis of 280 Research Papers. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 748–768.

[33] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *WSDM*.

[34] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[35] Megumi Ichimiya, Hope Muller-Tabanera, Jennifer Cantrell, Jeffrey B Bingenheimer, Raquel Gerard, Elizabeth C Hair, Dante Donati, Nandan Rao, and W Douglas Evans. 2023. Evaluation of response to incentive recruitment strategies in a social media-based survey. *Digital Health* (2023).

[36] Wooyoung William Jang and Kevin K Byon. 2020. Antecedents of esports gameplay intention: Genre as a moderator. *Comp. Human Behavior* (2020).

[37] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM CSUR* (2022).

[38] Linda K Kaye, Rachel Kowert, and Sally Quinn. 2017. The role of social identity and online social capital on psychosocial outcomes in MMO players. *Computers in Human Behavior* 74 (2017), 215–223.

[39] Jonathan W Kelly, Lucia A Cherep, Alex F Lim, Taylor Doty, and Stephen B Gilbert. 2021. Who are virtual reality headset owners? a survey and comparison of headset owners and non-owners. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 687–694.

[40] Daniel L King, Alex MT Russell, Paul H Delfabbro, and Dean Polisena. 2020. Fortnite microtransaction spending was associated with peers' purchasing behaviors but not gaming disorder symptoms. *Addictive Behaviors* 104 (2020), 106311.

[41] Rachel Kowert, Alexi Martel, and William B Swann. 2022. Not just a game: Identity fusion and extremism in gaming cultures. *Frontiers in Communication* 7 (2022), 1007128.

[42] Rachel Kowert and Julian A Oldmeadow. 2013. (A) Social reputation: Exploring the relationship between online video game involvement and social competence. *Computers in Human Behavior* 29, 4 (2013), 1872–1878.

[43] Andrey Krekhov, Sebastian Cmentowski, and Jens Krüger. 2019. The illusion of animal body ownership and its potential for virtual reality games. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[44] Jacob Leon Kröger, Philip Raschke, Jessica Percy Campbell, and Stefan Ullrich. 2023. Surveilling the gamers: Privacy impacts of the video game industry. *Entertainment Computing* 44 (2023), 100537.

[45] Chanel J Larche, Katrina Chini, Christopher Lee, and Mike J Dixon. 2023. To pay or just play? Examining individual differences between purchasers and earners of loot boxes in overwatch. *Journal of Gambling Studies* 39, 2 (2023), 625–643.

[46] Pascal Lessel, Maximilian Altmeyer, Julian Sahner, and Antonio Krüger. 2022. Streamer's Hell-Investigating Audience Influence in Live-Streams Beyond the Game. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY (2022), 1–27.

[47] Wen Li Anthony, Devin J Mills, Jackie F Stanmyre, and Lia Nower. 2023. Facets of mindfulness among video game players: A latent profile analysis. *Psychology of Addictive Behaviors* (2023).

[48] Jian Lu, Wei Li, Qingren Wang, and Yiwen Zhang. 2020. Research on data quality control of crowdsourcing annotation: A survey. In *IEEE DASC/PiCom/CBDCom/CyberSciTech*.

[49] Alexandra Sasha Luccioni and David Rolnick. 2023. Bugs in the data: How ImageNet misrepresents biodiversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14382–14390.

[50] Catherine C Marshall, Partha SR Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M Shipman. 2023. Who broke Amazon Mechanical Turk? An analysis of crowdsourcing data quality over time. In *Proceedings of the 15th ACM Web Science Conference 2023*. 335–345.

[51] Tara McAllister Byun, Peter F. Halpin, and Daniel Szeredi. 2015. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders* (2015).

[52] Francesco Neri, Carmelo Luca Smeralda, Davide Momi, Giulia Sprugnoli, Arianna Menardi, Salvatore Ferrone, Simone Rossi, Alessandro Rossi, Giorgio Di Lorenzo, and Emiliano Santarnecchi. 2021. Personalized adaptive training improves performance at a professional first-person shooter action videogame. *Frontiers in Psychology* 12 (2021), 598410.

[53] US Department of the Treasury. 2023. Sensitive Personal Data. https://web.archive.org/web/20231130015633/https://home.treasury.gov/taxonomy/term/7651.

[54] College of Wooster Human Subject Research Committee. 2023. *Guidance for Using MTurk*. https://web.archive.org/web/20230329032500/https://inside.wooster.edu/hsrc/guidance-for-using-mturk/

[55] Vincent Huard Pelletier, Arianne Lessard, Florence Piché, Charles Tétreau, and Martin Descarreaux. 2020. Video games and their associations with physical health: A scoping review. *BMJ open sport & exercise medicine* 6, 1 (2020), e000832.

[56] Cristiano Politowski, Fabio Petrillo, and Yann-Gaël Guéhéneuc. 2021. A survey of video game testing. In *IEEE/ACM AST*.

[57] Prolific. 2024. What is your pricing? researcher-help.prolific.com/hc/en-gb/articles/360009223533-What-is-your-pricing. Accessed in May 2024.

[58] Lisa Raith, Julie Bignill, Vasileios Stavropoulos, Prudence Millear, Andrew Allen, Helen M Stallman, Jonathan Mason, Tamara De Regt, Andrew Wood, and Lee Kannis-Dymand. 2021. Massively multiplayer online games and well-being: A systematic literature review. *Frontiers in Psychology* 12 (2021), 698799.

[59] Charles Reynaldo, Ryan Christian, Hansel Hosea, and Alexander AS Gunawan. 2021. Using video games to improve capabilities in decision making and cognitive skill: a literature review. *Procedia Computer Science* 179 (2021), 211–221.

[60] Jonathan Robinson, Cheskie Rosenzweig, Aaron J Moss, and Leib Litman. 2019. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PloS one* 14, 12 (2019), e0226394.

[61] Antonios Saravanos, Stavros Zervoudakis, Dongnanzi Zheng, Neil Stott, Bohdan Hawryluk, and Donatella Delfino. 2021. The hidden cost of using Amazon Mechanical Turk for research. In *HCI*.

[62] Johannes Schneider and Giovanni Apruzzese. 2023. Dual adversarial attacks: Fooling humans and classifiers. *JISA* (2023).

[63] Peter G Schrader, Mark C Carroll, Michael P McCreery, and Danielle L Head. 2020. Mixed methods for human–computer interactions research: An iterative study using Reddit and social media. *Journal of Educational Computing Research* 58, 4 (2020), 818–841.

[64] David Sharek and Eric Wiebe. 2014. Measuring video game engagement through the cognitive and affective dimensions. *Simulation & Gaming* 45, 4-5 (2014), 569–592.

[65] Zhiguo Shi, Guang Yang, Xiaowen Gong, Shibo He, and Jiming Chen. 2021. Quality-aware incentive mechanisms under social influences in data crowdsourcing. *IEEE/ACM Transactions on Networking* 30, 1 (2021), 176–189.

[66] Pier Paolo Tricomi, Lisa Facciolo, Giovanni Apruzzese, and Mauro Conti. 2023. Attribute inference attacks in online multiplayer video games: A case study on Dota2. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*. 27–38.

[67] Kellie Vella, Daniel Johnson, Vanessa Wan Sze Cheng, Tracey Davenport, Jo Mitchell, Madison Klarkowski, and Cody Phillips. 2019. A sense of belonging: Pokémon GO and social connectedness. *Games and Culture* 14, 6 (2019), 583–603.

[68] Zhao Wang, Anna Sapienza, Aron Culotta, and Emilio Ferrara. 2019. Personality and behavior in role-based online games. In *IEEE CoG*.

[69] Kerri Wazny. 2017. "Crowdsourcing" ten years in: A review. *Journal of global health* 7, 2 (2017).

[70] Donghee Yvette Wohn, Rabindra Ratan, and Leticia Cherchiglia. 2020. Gender and Genre Differences in Multiplayer Gaming Motivations. In *HCI in Games*.

[71] Lin Zhu. 2021. The psychology behind video games during COVID-19 pandemic: A case study of Animal Crossing: New Horizons. *Human Behavior and Emerging Technologies* 3, 1 (2021), 157–159.