Getting What You Paid For: Assessing Participant Experience Parameters for Amazon Mechanical Turk (MTurk) Workers on Survey Response Quality

M. Kumar*, D. Kim*, P.A. Weisgarber*, J.S. Valacich* and J.L. Jenkins**
* University of Arizona/Management Information Systems Department, Tucson, AZ, USA
** Brigham Young University/Information Systems Department, Provo, UT, USA
manasvik@arizona.edu

***Abstract* – MTurk is a powerful and widely used method for completing online Human Intelligence Tasks (HITs) such as website testing or completing psychological surveys. Researchers typically use various platform-provided work experience parameters to choose qualified and unqualified participants. Conventional wisdom suggests that participants with greater experience and historically high acceptance rates (i.e., higher-rated) will generate better quality data than their peers with lower experience and low acceptance rates (i.e., lower-rated). We examine the limits of this assumption by comparing responses and engagement behaviors between higher-rated, experienced (HE) participants and lower-rated, inexperienced (LI) participants while answering online surveys. We administered an online survey where participants first answered questions related to their MTurk account profile and then answered questions related to their personality. LI participants provide more inaccurate responses when answering factual questions (i.e., higher error rates) compared to HE participants. They also exhibit lower engagement behaviors when answering personality-related questions, resulting in marginally lower reliability scores for survey constructs. We are encouraged by the systematic differences we observed between the two populations and urge researchers to consider our findings before selecting optimal work experience parameters.**

Keywords – MTurk, Online Surveys, Data Quality

I. Introduction

Online surveys are a popular method utilized by researchers to collect data. While other alternatives exist [1], Amazon Mechanical Turk (MTurk) remains the most popular crowdsourcing platform that enables requesters (researchers and practitioners) to post several Human Intelligence Tasks (HITs) that workers can complete. Reference [2] suggests that compared to other traditional sample sources (i.e., university students), MTurk remains the most widely used platform due to its vast participant pool, convenient data collection, economical cost, and flexible research design options. Despite its popularity, MTurk faces several challenges due to its small, active subset of workers who complete most of the HITs available. It is estimated that nearly 80% of all HITs are completed by 20% of its participant pool [3]. Therefore, there are concerns that the platform is ineffective as the few workers who complete most of the tasks are familiar with the traditional techniques used to ensure data quality and can successfully circumvent them [4].

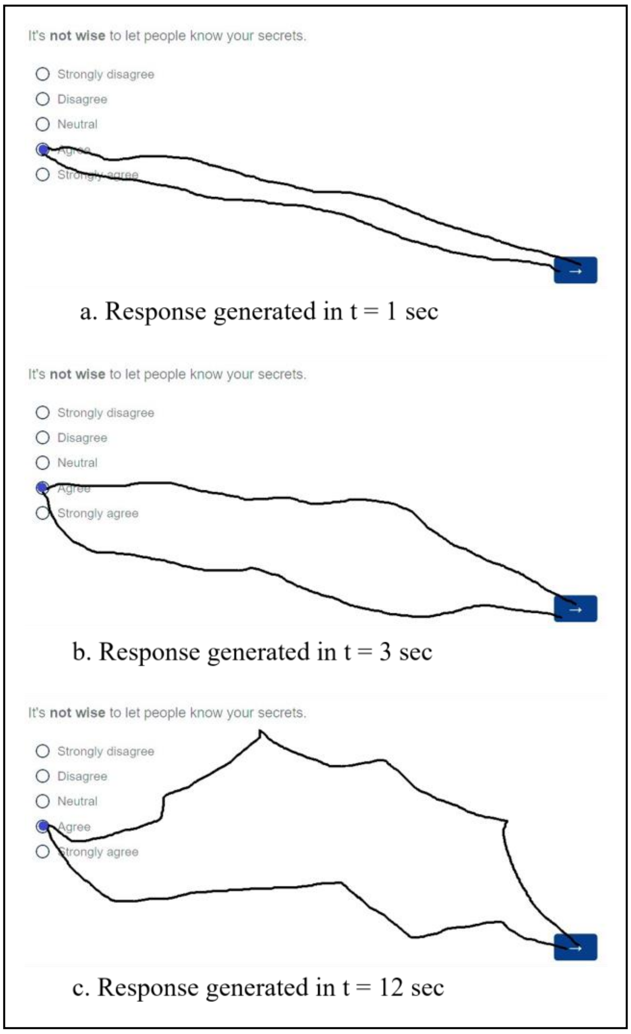
To combat these participant pool selection challenges, MTurk provides certain workers with the qualification of a "Master Worker": a qualification that is provided after the workers submit high-quality work for a sustained period. While the explicit details of how a worker becomes a "Master Worker" are not shared, they are often more expensive and do not provide significantly better data than other workers [5]. MTurk also allows its requesters to screen participants for its HITs based on their HIT acceptance rate, number of HITs completed or other criteria to ensure that adequate data quality requirements may be satisfied, and expenses are decided accordingly. In this study, we address where this separation line might be as we adjust the various qualification criteria to determine when higher-rated, experienced (HE) participants generate better quality data compared to lower-rated, inexperienced (LI) participants. While other studies have investigated the quality of data generated from less commonly used sections of MTurk's participant pool using survey responses alone [6], our study utilizes both response and engagement behavior to compare and evaluate the quality of data generated by participants of vastly different HIT acceptance and experience levels. By unobtrusively capturing paradata [7] (i.e., data such as mouse movements, clicks, timestamps etc., that is generated during the response generation process), we utilize fine-grained movement-based metrics to identify potentially problematic response generation behaviors when a participant answers a survey question. Thus, by using varied data sources, we provide a holistic comparison of the differences between data generated by HE and LI participants.

To demonstrate the efficacy of our approach, we administered a survey where HE and LI participants completed a two-part online survey where they answered factual questions (i.e., questions about their account) and traditional self-report survey questions (i.e., questions about their personality). Our analysis indicates that there are distinctive differences between HE and LI participants both in terms of their responses and their engagement behaviors. While answering factual questions, HE participants typically have lower error rates compared to their LI peers. They also spend more time generating the response to these questions. While answering traditional survey questions, LI participants generate responses that have marginally lower reliability scores and they often engage in extreme response generation behavior by requiring greater response times. Our results suggest that HE participants generate marginally better quality data than LI participants, suggesting that conventional wisdom on utilizing experience and acceptance rates to select participants is justified. The rest of the paper is organized as follows. First, we will briefly review the concerns regarding data quality in online surveys and how paradata may be used to address them. Next, we briefly describe the various aspects of the methodological approach. We will then present the results from our study, followed by a brief discussion and the conclusion.

II. Background

Data collected using online surveys may contain responses that are "invalid" (i.e., does not represent the true response value for a given question) and therefore poor quality [8]. A family of techniques commonly used to identify such responses involves the use of attention check questions. These questions require the participant to select a specific response (e.g., "Please select Moderately Inaccurate for this item."), as discussed in [9]. As mentioned earlier, one of the biggest concerns associated with MTurk participants is the small percentage of active workers who complete a large proportion of HITs. As a result of these numerous interactions, these workers become skilled in identifying attention check questions easily [10], thereby negating one of the most popular approaches used by researchers to identify inattentive participants who generate such invalid responses. Therefore, current traditional approaches may not be adequate to identify inattentive participants, especially if they may have answered similar questions earlier.

An alternative approach to identifying responses generated by inattentive participants involves collecting paradata. To illustrate the effectiveness of this approach, consider the three hypothetical responses to a survey question depicted in Figure 1. In Parts a, b, and c, the participants chose to select the same option: "Agree". However, the time taken to generate the responses (i.e., response time) varies greatly. Without using paradata, it appears that all three participants generated the same response, and it is impossible to distinguish between them. In contrast, by observing the response times, it is now possible to comment on the response generation process of each participant and therefore on the true "accuracy" of each response. Anecdotally, it has been observed that minimum amount of time to generate a meaningful response is 2 seconds [9]. This contrasts with the time taken by participant A, who takes around 1 second to generate a response. It is more likely that this participant didn't pay adequate attention while generating the response for this question, and therefore, this response is more likely to be invalid. Note that this response is also characterized by decisive, quick movements with less deviations. Participant B takes around 3 seconds to generate a response. This is close to the median response time taken by other participants to answer this question. As the participant had sufficient time to read the question and the movement is characterized by adequate, precise, deliberation; it is more likely that the response generated is a good quality response. Participant C takes 12 seconds to generate their response. This is significantly longer than the time taken by most participants to answer this question. Compared to previous participants, the movement here is characterized by greater deviations, potential pauses, and indecisive movements. It is therefore more likely that the response generated is invalid, as the participant may have been distracted or multitasking. Therefore, by utilizing fine-grained mouse cursor movement, it is possible to derive metrics (i.e., response time) that can be used to understand engagement behavior at an item level. These metrics may also be used to identify particularly poor-quality responses, or bad data, in a manner that is completely unobtrusive and can be performed after data collection has been completed.



It's **not wise** to let people know your secrets.

○ Strongly disagree
○ Disagree
○ Neutral
● Agree
○ Strongly agree

a. Response generated in t = 1 sec

It's **not wise** to let people know your secrets.

○ Strongly disagree
○ Disagree
○ Neutral
● Agree
○ Strongly agree

b. Response generated in t = 3 sec

It's **not wise** to let people know your secrets.

○ Strongly disagree
○ Disagree
○ Neutral
● Agree
○ Strongly agree

c. Response generated in t = 12 sec

Figure 1: Representation of responses generated by three hypothetical participants – t = Response times with mouse movement navigation.

Some of the major challenges of MTurk research include inattention, misrepresentation, selection bias, non-naivete and social-desirability bias [2]. As illustrated earlier, paradata may be used to improve data quality by capturing engagement behaviors during response generation. Engagement behavior metrics have been used to examine a variety of phenomena, including multitasking [11], social desirability bias [12], distractions [13], adequacy of outlier definitions [14], and impact of question formats on response quality [15]. In particular, there is comprehensive research that examines the impact of using various response time-based metrics to improve response quality [16]. In this study, we utilize both survey responses and engagement behaviors to compare data quality of responses generated by HE and LI participants.

III. Methodology

We designed a two-part study to compare responses and engagement behaviors between HE and LI participants. In the first part, we asked participants to select information about their MTurk account, including information about the number of HITs they had completed and their HIT approval rating. In the second part of the study, they self-report scores about their personality dimensions of psychopathy and sadism. Data collection occurred in two phases. In the first phase, we collected data from LI participants while in the second phase, we collected data from HE participants. Therefore, both groups of participants (HE and LI participants) answered factual questions (i.e., questions about their account) and traditional self-report survey questions (i.e., questions about their personality). We also captured fine-grained mouse cursor data while the participants generated their responses. Thus, the data collected from this study allows us to compare both responses and engagement behaviors between HE and LI participants.

A. *Pilot Studies*

Prior to conducting the primary study, we conducted a series of pilot studies to determine optimal thresholds for experience and HIT acceptance rates in MTurk participants. As we were interested in comparing workers who hadn't achieved "Master" status, we considered different thresholds to adequately represent typical experience and acceptance rates of other "regular" workers. Reference [5] found that the average number of HITs completed by a regular worker is around 19,791 while the average approval rate is around 97.9%. We concluded that a LI participant should have a lower threshold compared to the average participant; therefore we selected appropriate thresholds for experience (i.e., 5,000 and 10,000) and acceptance rates (i.e., 90 and 95). We then conducted several pilot studies to determine the availability of participants satisfying these threshold buckets. We found that to collect a typical number of participants (n=100), it is advisable to select a threshold of 10,000 for HITs completed and 90 to 95 for acceptance rate. Therefore, we determined that participants who completed lesser than 10,000 HITs and have an HIT acceptance rate between 90 to 95 are LI participants. Consequently, participants who completed greater than 10,000 HITs and have an HIT acceptance rate of 95 and above are HE participants.

B. *Survey Design*

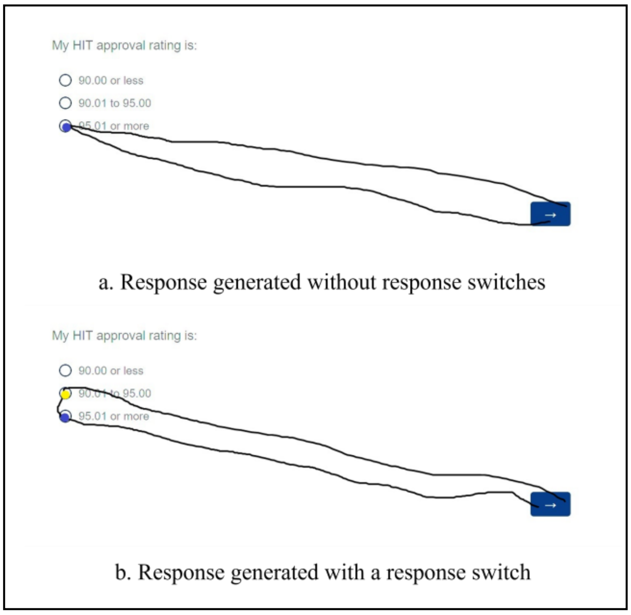
In the first part of the two-part survey, participants were required to provide information about their MTurk account. This section includes information about their HIT approval rating (categorized as 90 or less, 90 to 95, and 95 or more) and the number of HITs they had completed (categorized as 0 to 10,000 and 10,001 or more), in accordance with the thresholds determined in the pilot studies. The participants were requested to answer these questions "as best as they can", and that these responses may be used to "recruit them in future studies." We also included an attention question ("22 times 22 is:") to identify inattentive participants. Only participants who successfully answered the attention check question could participate in the second part of the study. In the second part of the study, participants were required to provide self-report responses to the Dark-Tetrad Personality questionnaire, as adapted from [17]. Items for all measures utilized a 5-point Likert Scale ranging from "Strongly Disagree" to "Strongly Agree."

C. *Participants*

We recruited participants from Amazon Mechanical Turk in two phases. In the first phase, participants who completed less than 10,000 HITs and had a HIT approval rating lower than 95% (LI participants) were recruited for the study. In the second phase, participants who completed greater than 10,000 HITs and had a HIT approval rating higher than 95% were recruited (HE participants). The study consisted of two parts, and participants were paid \$0.05 for completing the first part and \$0.95 for completing the second part of the study. Only data from participants who completed both parts of the study were used in future analyses. Of the 200 participants recruited for the study (100 for each phase), 84 participants completed both parts of the study in the first phase, while 91 participants completed both parts of the study in the second phase.

D. *Engagement Data Tracking*

We utilized the Qualtrics survey system to host the survey used for data collection. A custom JavaScript library that captures mouse-cursor movements (e.g., x and y coordinates of the cursor, timestamps, etc.) and other behavioral data (e.g., mouse clicks, interaction with HTML elements, etc.) was embedded in the survey to track engagement behavior. The data captured by this library was sent to a web service for storage and further processing. Through in-house python code developed by the research team, we create specialized metrics to understand engagement behaviors of participants at a question level. In this study, we utilize two metrics: response time and response switches. Response time may be understood as the time taken by the participant to generate a response for a survey question. Response switches indicate the number of intermediate selections made prior to selecting a final response. These metrics have previously been shown to identify poor quality responses in online surveys [18]. For example, excessive response switches could indicate indecision, lack of comprehension, or the presence of response bias. Consider Figure 2, which presents two responses to the question of HIT approval ratings used in this study. In Part a, the participant answers the question directly without any intermediate selections. However, Part b shows the response generated by the participant, who has a response



a. Response generated without response switches

b. Response generated with a response switch

Figure 2: Representation of responses generated by two hypothetical participants – Response Switches with mouse movement navigation.

switch. It is likely that this participant chose to select a lower approval rating before finally deciding to select the alternative. As mentioned earlier, this response behavior could indicate indecision, lack of comprehension, or (as is more likely in this case) the presence of a social response bias. Similarly, while response times have been used as a proxy for engagement, extreme response generation behaviors: either too fast or too slow, may not be ideal. To determine these engagement behavioral "outliers", we adopted a procedure similar to that utilized in [18]. We first compute the average response time utilized by an individual to generate a response for questions pertaining to a particular construct. We then compute the median (M) and median absolute deviation (MAD) times of all average response times computed for all individuals. If an individual's average response time is greater than M + 3*MAD, then the participant spent an unusual amount of time answering questions pertaining to the construct and is therefore considered an engagement behavioral outlier.

IV. Analysis and Results

In this section, we provide results for the various analyses conducted to compare responses and engagement behaviors between HE and LI participants. As mentioned earlier, both groups of participants provided responses to factual questions (i.e., questions about their account) and traditional self-report survey questions (i.e., questions about their personality). In addition, we also collected fine-grained mouse movement data while these responses were generated using a JavaScript library. Previous studies have shown that relevant metrics derived from such data may be used to identify responses of poor data quality as they provide information about participant engagement behaviors [19]. In this study, we answer the following questions:

- How do responses obtained by HE and LI participants differ for factual questions?
- How do responses obtained by HE and LI participants differ for traditional survey questions?
- How does engagement behavior exhibited by HE and LI participants differ for factual questions?
- How does engagement behavior exhibited by HE and LI participants differ for traditional survey questions?

A. *Response Data*

We compared the survey responses received from both HE and LI participants for different question types (i.e., factual questions and traditional self-report survey questions). In factual questions, we asked participants to select the number of HITs they completed and their HIT approval rate "as best as they can." As we know the actual values of these parameters, we determined the error rate of their responses. For the HITs completed question, 5% of HE participants provided erroneous answers while 8% of LI participants provided erroneous answers. For the HIT approval rate question, the difference is more apparent. 49% of LI participants provided erroneous answers for this question, while around 6% of HE participants provided erroneous answers. Thus, for both factual questions, HE participants provided less erroneous answers compared to LI participants.

We also compared the survey responses of HE and LI participants to more traditional survey questions (i.e., the Dark Tetrad Questionnaire). One approach to evaluate the quality of survey data is through reliability, with coefficients greater than 0.7 being the recommended standard for MTurk data [20]. For our study, we evaluate the Cronbach's alpha value for the constructs of psychopathy and sadism for data from HE and LI separately. For psychopathy, the Cronbach's alpha obtained from analyzing HE and LI participants' survey responses is 0.931 and 0.886, respectively. Similarly, the Cronbach's alpha obtained by analyzing the responses from HE and LI participants for sadism is 0.89 and 0.843, respectively. In either case, the values of Cronbach's alpha are greater for HE participants than LI participants, while both are above the recommended threshold.

B. *Engagement Data*

We compared engagement behaviors through metrics generated from fine-grained mouse movement data as the participants generated their responses to both factual questions and traditional self-report survey questions. For the factual questions, we created several question-level metrics and analyzed for differences between HE and LI participants. We found that when participants are asked to report the number of HITs they've completed, the HE participants have greater response times compared to their LI peers (t(114)=-2.87, p<0.01). Note that HE participants generated fewer erroneous responses on this question compared to the LI participants. Similarly, we found that when participants are asked to report their HIT approval rating, LI participants have a significantly greater number of response switches compared to their HE peers (t(119)=2.68, p<0.01). It is important to note that HE participants generated significantly fewer erroneous responses compared to LI participants. The implications of these results are discussed in the next section.

TABLE I. SUMMARIZED RESULTS OF ALL ANALYSES PERFORMED

Type size	**Response**		**Engagement Behavior**	
	HE Participant	***LI Participant***	***HE Participant***	***LI Participant***
HIT Completion Question	Error (5%)	Error (8%)	Greater Response Time	Less Response Time
HIT Approval Question	Error (6%)	Error (49%)	Fewer Response Switches	More Response Switches
Survey Questions	Cronbach's Alpha (0.931 and 0.89)	Cronbach's Alpha (0.886 and 0.843)	Less proportion of slow-moving outliers	Greater proportion of slow-moving outliers

We analyzed the difference in engagement behaviors for the traditional self-report survey questions by identifying engagement behavioral outliers, as discussed in the Methodology section. For questions pertaining to psychopathy construct, we found that the participants who exhibited extreme engagement behaviors (greater response times) are disproportionately LI participants (52% compared to a baseline of 48%). We observe similar results when analyzing extreme engagement behaviors for questions pertaining to sadism construct as well (53% compared to a baseline of 48%). Thus, a greater proportion of engagement behavioral outliers were found to be LI participants for both constructs. Table 1. summarizes the results obtained for all analyses performed.

V. DISCUSSION

While comparing differences in responses between HE and LI participants, the results obtained are intuitive. For the factual questions, we observed that HE participants produced less erroneous responses compared to LI participants. This is expected as HE participants are more careful and aware of important questions in the survey compared to LI participants. It is also worth mentioning that the error rate of LI participants on the HIT approval rate question (49%) is significantly greater than the error rates among others. This is in part due to the instruction provided which indicated that their responses may be used for future recruitment. It is possible that the LI participants are aware of the importance researchers place on HIT approval rates in their selection process and decided to be deceptive to ensure future selection. Comparing differences in responses to the traditional survey questions revealed similarly intuitive results. The Cronbach's alpha value in all cases is above the recommended threshold, as expected for a widely used survey. While the Cronbach's alpha values are greater for responses generated by HE participants, the differences are not large enough to warrant further analysis.

Comparing differences in engagement behaviors between HE and LI participants provides several interesting results, especially when viewed in conjunction with differences in responses. For the factual question concerning number of HITs completed, we observed that HE participants have significantly greater response times than LI participants. Note that for this question, the error rate among HE participants is lower. This is intuitive, as response times are often seen as a proxy for engagement, and it is likely that HE participants spent more time on this question to ensure that they answered it correctly. For the other factual question concerning the HIT approval rate, we observed that LI participants have significantly greater number of response switches compared to HI participants. They also have a significantly large error rate on this question. As mentioned earlier, a large number of response switches indicates indecision, improper comprehension or the presence of response bias. In this study, it was clearly indicated that participants should answer the questions "as best as they can" and that their responses may be used for future recruitment. It is more likely that the LI participants engaged in social desirability bias and selected options that inflated their approval ratings.

Comparing engagement behaviors between HE and LI participants while answering traditional survey questions revealed that LI participants are likely to have engaged in extreme engagement behaviors (i.e., have unusually greater response times). As mentioned earlier, participants who engage in extreme response generation behaviors are found to generate poor quality responses. Note that for these questions, the Cronbach's alpha (while above the recommended value) is lower for responses generated by LI participants. It is typically assumed that relatively inexperienced "amateurs" are easily distracted from their work [10], and these distractions could result in excessive response times among these participants.

In addition to highlighting the response and behavioral differences between HE and LI participants while answering survey questions, we also demonstrate the effectiveness of fine-grained mouse movement-based metrics to identify poor quality responses. While several studies have identified and acknowledged concerns with MTurk's participant pool [21], MTurk (and other related crowdsourcing services) provides researchers with access to several features that can mitigate common concerns [22]. By integrating a custom JavaScript code to collect data and developing useful metrics to understand engagement behavior, we provide researchers with an alternative approach to identifying poor quality responses that goes beyond existing practices. The suggested approach may also be used on other crowdsourcing services or traditional recruitment approaches that lack prescreening and other related features provided by MTurk to maintain data quality standards.

A. Limitations

In this study, participants were required to answer factual questions about their MTurk account. They were instructed to answer these questions "as best as they can", and that their responses may be used for future recruitments. While these instructions were intended to ensure correctness of their response, the LI participants may have intentionally provided deceptive responses to improve their odds of recruitment. This wasn't possible for the HE participants as they couldn't provide a more deceptive, inaccurate answer. In the future, we advise researchers to utilize factual questions that do not provide any group of participants unintended incentives to provide inaccurate responses.

In this study, we compared responses and response generation behaviors between HE and LI participants. We could not collect data for higher-rated, inexperienced participants and lower-rated, experienced participants as these groups of participants did not have easily available workers. It is, however, possible that by utilizing a different threshold or multiple thresholds (i.e., experienced > 100,000 HITs and inexperienced < 10,000 HITs), we may obtain a set of available workers whose data may be compared. We acknowledge that these thresholds will continue to evolve as MTurk grows and advise researchers to utilize the latest available data to select appropriate thresholds.

VI. Conclusion

In this study, we examined whether traditionally higher-rated, experienced participants generate better quality data in online surveys compared to lower-rated, inexperienced participants. Both HE and LI participants were required to answer a survey containing factual and typical survey questions. We analyzed differences between both groups of participants not only through their responses but also using their engagement behaviors, as captured by fine-grained mouse movement-based metrics. Our analysis revealed that HE participants provide less erroneous responses to factual questions, likely because they spend more time on these questions and do not exhibit potentially deceptive behavior. We also found that LI participants provide slightly less reliable (but acceptable) responses to traditional survey questions and are more likely to engage in extreme engagement behaviors.

References

[1] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," in Journal of Experimental Social Psychology, 70, 2017, pp. 153-163.

[2] H. Aguinis, I. Villamor, and R. S. Ramani, "MTurk research: Review and recommendations," in Journal of Management 47, no. 4, 2021, pp. 823-837.

[3] K. Fort, G. Adda, and K. B. Cohen, "Amazon Mechanical Turk: Gold mine or coal mine?," in Computational Linguistics, 2011, pp. 413-420.

[4] A. W. Meade, and S. B. Craig, "Identifying careless responses in survey data," in Psychological methods 17, no. 3, 2012, pp. 437.

[5] E. Loepp, and J. T. Kelly, "Distinction without a difference? An assessment of MTurk Worker types," in Research & Politics 7, no. 1 , 2020, 2053168019901185.

[6] J. Robinson, C. Rosenzweig, A. J. Moss, and L. Litman, "Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool," in PloS one 14, no. 12 , 2019, e0226394.

[7] F. Kreuter, Improving surveys with paradata: Analytic uses of process information. John Wiley & Sons, 2013.

[8] P.G. Curran, "Methods for the detection of carelessly invalid responses in survey data," in Journal of Experimental Social Psychology, 66, 2016, pp. 4-19.

[9] J.L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon, "Detecting and deterring insufficient effort responding to surveys," in Journal of Business and Psychology 27 , 2012, pp. 99-114.

[10] E. Peer, J. Vosgerau, and A. Acquisti. "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk," in Behavior research methods 46 ,2014, pp. 1023-1031.

[11] L. Zwarun, and A. Hall. "What's going on? Age, distraction, and multitasking during online survey taking," in Computers in human behavior 41 , 2014, pp. 236-244.

[12] J.L. Jenkins, J. S. Valacich, and P. Williams, "Human-computer interaction movement indicators of response biases in online surveys," in International Conference on Information Systems, 2018, Seoul, Korea.

[13] A. Wenz, "Do distractions during web survey completion affect data quality? Findings from a laboratory experiment," in Social Science Computer Review 39, 2021, pp.148-161.

[14] J.K. Höhne, and S. Schlosser, "Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata," in Social Science Computer Review, 36, 2018, pp. 369-378.

[15] J.K. Höhne, S. Schlosser, and D. Krebs, "Investigating cognitive effort and response quality of question formats in web surveys using paradata," in Field Methods 29, 2017, pp. 365-382.

[16] M. Matjašič, V. Vehovar, and K.L. Manfreda, "Web survey paradata on response time outliers: A systematic literature review," in Advances in Methodology and Statistics, 15, 2018, pp. 23-41.

[17] D.L. Paulhus, "Toward a taxonomy of dark personalities," in Current Directions in Psychological Science, 23, no. 6 , 2014, pp. 421-426.

[18] M. Kumar, D. Kim, , J. S. Valacich, J. L. Jenkins, and A. Dennis, "Improving the Quality of Survey Data: Using Answering Behavior as an Alternative Method for Detecting Biased Respondents," 2021, SIGHCI 2021 Proceedings, 13.

[19] M. Kumar, D. Kim, , J. S. Valacich, and J. L. Jenkins, “Too Fast? Too Slow? A Novel Approach for Identifying Extreme Response Behavior in Online Surveys,” 2022, SIGHCI 2022 Proceedings, 14.

[20] C.J. Holden, T. Dennie, and A. D. Hicks, "Assessing the reliability of the M5-120 on Amazon's Mechanical Turk," in Computers in Human Behavior 29, no. 4 , 2013, pp. 1749-1754.

[21] D. Hauser, G. Paolacci, and J. Chandler, “Common concerns with MTurk as a participant pool: Evidence and solutions,” in F. R. Kardes, P. M. Herr, and N. Schwarz (Eds.), Handbook of research methods in consumer psychology, 2019, pp. 319–337

[22] J. Chandler, P. Mueller, and G. Paolacci, “Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers,” Behavior research methods, 46, 2014, pp. 112-130.