Transactions on Replication Research

Methodological Replication DOI: 10.17705/1atrr.00082 ISSN 2473-3458

# Mechanical Turk Versus Student Samples: Comparisons and Recommendations

**Stephen A. De Lurgio II**
Department of Information Systems, Walton College,
University of Arkansas
*sadelurg@uark.edu*

**Amber Young**
Department of Information Systems, Walton College,
University of Arkansas
*agy002@uark.edu*

**Zachary R. Steelman**
Department of Information Systems, Walton College,
University of Arkansas
*zsteelma@uark.edu*

**Abstract:**

Mechanical Turk and other online crowdsourcing markets (OCMs) have become a go-to data source across scientific disciplines. In 2014 Steelman and colleagues investigated how Mechanical Turk data compared with student samples and consumer panels. They found the data to be comparable and reliable for academic research. In the nearly 10 years since its publication, the use of Mechanical Turk in research has grown substantially. To understand whether their results still hold, we conducted a partial replication to determine how Mechanical Turk workers continue to compare with students using UTAUT 2 as our theoretical model and virtual-reality headsets as the focal IT artifact. Our findings generally align with Steelman et al. (2014) and confirm that Mechanical Turk continues to offer a suitable alternative to student samples. This study reveals consistent results between the student and OCM samples, indicating the potential for interchangeability. The OCM samples are primarily male, while the student sample is majority female, following current US academic trends. All samples are significantly different in age, and only the US OCM and non-US OCM samples are similar in education. The path coefficients from the non-US OCM sample differ significantly from those from other OCM samples; the path coefficients derived from the student sample do not differ significantly from any OCM sample. While sample differences exist, as expected, many are addressable post hoc if anticipated and designed for during data collection. From our findings and the extant literature, we summarize recommendations for researchers and review teams.

**Keywords:** Survey, Crowdsourcing, MTurk, Mechanical Turk, Survey Research, Replication, Research Methods, Survey Methods, Empirical Research, Sampling, Questionnaire Surveys, Data Collection.

The manuscript was received 04/05/2022 and was with the authors 8 months for 3 revisions.

# 1 Introduction

Student samples are useful for studying topics like consumer behavior and technology use where differences between students and the general population are unlikely to be significant. Online crowdsourcing markets (OCM) like Mechanical Turk emerged in the early 2000s allowing researchers to collect anonymous data without relying on student samples. Steelman and his team compared data quality and validity in 2011 using students, consumer panels, and Mechanical Turk workers. Published by MIS Quarterly in 2014, they found that OCM samples from the United States "produced models that lead to similar statistical conclusions as both the US students and the US consumer panels at a considerably reduced cost" (Steelman et al., 2014, p. 355).

Many IS journals have published studies using OCM data, including MIS Quarterly (e.g., Jia et al., 2022), the Journal of MIS (e.g., Lowry et al., 2017), the Journal of the AIS (e.g., Jiang et al., 2021), the Journal of Strategic Information Systems (e.g., Lowry & Wilson, 2016), and others. This growth in OCM usage brought significant changes, including professionalizing microtask work and the use of bots. In response, some researchers and editors lost faith in OCM samples (Ford, 2017; Landers & Behrend, 2015). Many authors (ourselves included) experienced pushback from review teams that may have been unfamiliar or uncomfortable with OCM samples. Some journals in adjacent fields stopped accepting manuscripts utilizing OCMs (Landers & Behrend, 2015; Walter et al., 2019). Our study aims to determine if OCM data is still a valid option for researchers by comparing students and OCM samples. In addition to adding to the literature on the validity and use of OCM samples, this partial replication of Steelman et al. (2014) increases confidence in the applicability of their methods and findings (Dennis & Valacich, 2014).

## 1.1 Overview of Original Research

Steelman et al. (2014) compared survey results from US students, consumer panels, and Amazon's Mechanical Turk. They used the technology acceptance model (TAM) to explore differences in demographics and measurements. Their focus was not on testing the theoretical model itself, but on utilizing a well-accepted model to focus on the differences across their collected samples. Steelman et al. (2014)'s main takeaway was that OCM samples were a cost-effective alternative to student samples and consumer panels in IS research without compromising quality and validity.

## 1.2 Overview of this Replication

This study is a conceptual replication (Dennis & Valacich, 2014) of Steelman et al. (2014). We used similar methods, contexts, treatments, and analyses, with some modifications, to test the robustness of their findings. However, we made three changes: we replaced the TAM with the UTAUT 2 as the theoretical model, used the Oculus Quest 2 as the focal technology instead of Windows 7, and omitted the consumer panel. These changes ensured reliable and consistent comparisons with previous research. The UTAUT 2 adoption model is better suited to consumer contexts. The shift to the Oculus Quest 2 allowed participants to envision adopting a technology they were likely familiar with but had not yet adopted, similar to Windows 7 in the original study. Given the declining popularity of expensive consumer panels, we eschewed comparing them and instead focused on comparing student and OCM samples. We utilized CloudResearch's MTurk Toolkit[1] to implement Mechanical Turk (Litman et al., 2017). This toolkit provided more extensive criteria for filtering and managing participants than Mechanical Turk alone.

# 2 Methods

## 2.1 Participant Samples

We gathered data from students at a Midwestern university and workers on Amazon's Mechanical Turk. Like Steelman et al. (2014), we collected data from US-based participants, non-US-based participants, and worldwide participants from Mechanical Turk using CloudResearch's MTurk Toolkit. We utilized CloudResearch's MTurk Toolkit to manage participant compensation, screen by location, control numbers, and select based on HIT completion and approval rating. We used Mechanical Turk samples from workers with a 95% positive rating or higher (Jia et al., 2017; Peer et al., 2014; Steelman et al., 2014). We collected four samples with 2,281 responses and 1,612 remained after filtering.

[1] https://www.cloudresearch.com/products/turkprime-mturk-toolkit/

## 2.2 Survey Procedure

Our survey procedure followed Steelman et al. (2014). Participants provided consent, watched videos about the technology, and completed questionnaire items. The videos about Oculus Quest 2 were created by editing the official release videos[2]. The Quest 2 consists of hardware—a virtual-reality headset and two controllers—and software. These allow users to see, hear, speak, and interact with other users in various shared virtual-reality spaces (i.e., games, conference rooms, and design studios). The videos provided information for participants to determine their future adoption intentions; they detailed the system setup and five affordances provided by the platform: working in virtual reality, virtual meetings, virtual socializing, gaming, and exercising in virtual reality. After the videos, participants answered questions from the UTAUT 2 instrument and demographic questions. At the end of the survey, students provided their email addresses to receive class credit, and Mechanical Turk participants received payment codes.

The survey contained two attention-check questions to compare attentiveness rates among participants (Aust et al., 2013). One was embedded within another block of questions, and the other was isolated. All participants across all samples watched the same videos and received the same questionnaire.

OCM workers received $1 for completing surveys, while Steelman et al.'s (2014) participants received 20 cents. Recent studies suggest $1 or more for MTurk samples in IS (e.g., Jia et al., 2017). We excluded incomplete or rushed surveys. Students were not compensated, but some instructors offered extra credit.

## 2.3 Participant Removal

Following Steelman et al. (2014), we conducted minimal data cleaning while tracking differences across studies to compare data cleanliness and subsequent validity. Table 1 shows the criteria used in our minimal data-cleaning procedures and the number of records in each sample that met or failed the given criteria. Table 2 illustrates the data-cleaning steps, the number of records removed by sample group, and the remaining sample size following each step.

**Table 1. Screening Criteria by Sample**

| | Survey | | | | |
|---|---|---|---|---|---|
| | **Students** | **US OCM** | **Non-US OCM** | **Worldwide OCM** | **Total** |
| Collected | 172 | 696 | 690 | 718 | 2276 |
| Completed | 134 | 503 | 492 | 495 | 1,624 |
| **Passed Duration Check** | | | | | |
| No | 2 | 3 | 3 | 0 | 8 |
| Yes | 132 | 500 | 489 | 495 | 1616 |
| **Passed Isolated Attention Check** | | | | | |
| No | 3 | 4 | 2 | 0 | 9 |
| Yes | 131 | 499 | 490 | 495 | 1615 |
| **Passed Embedded Attention Check** | | | | | |
| No | 26 | 36 | 37 | 45 | 144 |
| Yes | 108 | 467 | 455 | 450 | 1480 |
| **Passed Price Check** | | | | | |
| No | 18 | 143 | 100 | 147 | 408 |
| Yes | 116 | 360 | 392 | 348 | 1216 |

We first removed records from participants who did not complete the survey. We observed a greater percentage of participants who started a survey and did not complete it than Steelman et al. (2014). Next, we removed records from participants who, as indicated by response time, likely had not completed watching the videos[3]. Then, we removed records from participants who did not pass the isolated attention-check question. Next, we removed participants who failed the isolated attention-check question but kept those who failed the more complex embedded attention- or price-check questions. The study compares minimally cleaned data, and non-obvious or non-coincident attention check failures do not necessarily

[2] https://www.oculus.com/quest-2/ (Facebook, n.d.)

[3] We chose eight minutes as a minimum because it is less than the total run time of the videos (08:20). Participants could increase playback speed (e.g., 1.5x, 2.0x, etc.). Still, it is unlikely that respondents who took less than eight minutes watched the videos. At the maximum speed of 2x, it would take four minutes and 10 seconds to watch the videos, leaving less than four minutes (3:50) to complete the rest of the survey.

indicate poor-quality responses (Hauser & Schwarz, 2015; Krosnick, 1991; Oppenheimer et al., 2009). Therefore, we retained participants who only failed the price check, embedded attention check, or both.

**Table 2. Participant Response Removal**

| | | Total Responses Collected | Step 1: Completion | Step 2: Passed Duration Check | Step 3: Passed Isolated Attention Check | Final Sample | Steelman et al. (2014, p. 361) |
|---|---|---|---|---|---|---|---|
| All samples | Sample | 2276 | 1624 | 1616 | 1612 | 1612 | 792 |
| | Dropped | 0 | -652 | -8 | -4 | -664 | -68 |
| | % Retained | 100% | 71% | 71% | 71% | 71% | 92% |
| Students | Sample | 172 | 134 | 132 | 131 | 131 | 165 |
| | Dropped | 0 | -38 | -2 | -1 | -41 | -13 |
| | % Retained | 100% | 78% | 77% | 76% | 76% | 93% |
| US OCM | Sample | 696 | 503 | 500 | 498 | 498 | 222 |
| | Dropped | 0 | -193 | -3 | -2 | -198 | -14 |
| | % Retained | 100% | 72% | 72% | 72% | 72% | 94% |
| Non-US OCM | Sample | 690 | 492 | 489 | 488 | 488 | 212 |
| | Dropped | 0 | -198 | -3 | -1 | -202 | -25 |
| | % Retained | 100% | 71% | 71% | 71% | 71% | 89% |
| Worldwide OCM | Sample | 718 | 495 | 495 | 495 | 495 | 193 |
| | Dropped | 0 | -223 | 0 | 0 | -223 | -16 |
| | % Retained | 100% | 69% | 69% | 69% | 69% | 92% |

Following Steelman et al. (2014), to understand differences and similarities across samples, we did not remove outliers or adjust the model, thereby retaining potentially divergent data instead of cleaning it out. The cleaning steps in Table 2 resulted in samples containing 131 student records, 495 worldwide OCM worker records, 488 non-US OCM worker records, and 498 US-based OCM worker records. While the raw number of dropped records differed across samples, their ratios were similar across all samples and steps ($\chi^2_{(9)} = 1.30$, $p = .998$), indicating that, to this point, the OCM samples do not require additional cleaning beyond a traditional student-convenience sample. Note that the results in Table 2 do not match those in Table 1. The criteria in Table 1 are not mutually exclusive, and the numbers in Table 2 depend on the order of the cleaning steps applied. Additionally, we note that, were this an actual acceptance and use study, further cleaning steps would be required.

# 3 Analyses and Results

We present our analysis and results (Table 3) in the same order as Steelman et al. (2014). We begin by discussing the demographic breakdown and differences of each sample. Then, we discuss the structural equation modeling (SEM) results. Finally, we address measurement invariance. Also, note the use of ***bold italicized*** text in the following tables. We use bold and italicized text to indicate statistically significant results or those outside an established threshold.

## 3.1 Demographics

First, we compared demographic differences across groups, shown in Table 4 and Table 5. All OCM samples had more males than nonmales. The worldwide OCM and non-US OCM samples had the same gender proportions as the US OCM sample, while the non-US OCM sample had a significantly larger proportion of males than the worldwide OCM. The gender proportions of the student sample (47.3% male) were significantly different from all the OCM samples.

All samples' mean ages were significantly different. US OCM sample participants are the oldest on average (~40 years old), and the student sample participants were the youngest on average (~22 years old). The student sample also had the smallest range of ages (19 to 54), and the worldwide OCM sample had the broadest (19 to 85).

All sample populations had significantly different mean levels of education except for the US OCM and non-US OCM samples. Participants in the US OCM and non-US OCM samples had the highest mean levels of education, followed by participants in the worldwide OCM sample. The student sample participants had the lowest mean level of education. The mean differences between the student and OCM

samples were all highly significant, and the worldwide OCM sample differed from both the US OCM and non-US OCM samples. We saw the same levels of significance repeated for the Wilcoxon rank-sum *z*-stat for the education proportions.

**Table 3. Analysis Procedures**

| Analysis Step | Focus Of Test | Empirical Tests | |
|---|---|---|---|
| | **Both Studies** | **Steelman et al. (2014) p.326** | **Current Study[4]** |
| **Demographics** | Differences in sample composition. | Chi-square proportion, *t*-tests, Wilcoxon Sum-rank. | Chi-square proportion, *t*-tests, Wilcoxon Sum-rank. |
| **Factor Analysis+** | Differences between samples regarding validity & reliability. | PCA: MLA w/Varimax and Oblimin rotation. Lambda values, CFI, SRMR RMSEA, Cronbach's alpha, composite reliability, reliability coefficient, AVE, Fornell-Larcker | Cronbach's alpha, composite reliability, reliability coefficient, AVE, Fornell-Larcker, Loadings-Crossloadings. |
| **Structural Model+** | Differences in construct relationships and pat1h coefficients. | CB-SEM, PLS-SEM, CFI, SRMR, RMSEA, $R^{2,}$ and *t*-tests. | PLS-SEM, $R^2$ and *t*-tests |
| **Measurement Invariance Tests** | Differences between samples in the configuration, composition, mean, and variance of the model constructs across samples. | Differences in sample intercepts (means), loadings, variances, ANOVAS, & pairwise comparisons. | MICOM procedure: correlation of composite score weights, differences in mean construct scores, and differences in the variance of construct scores across samples. |
| Notes: + Steelman et al. (2014) performed analyses using covariance-based and variance-based SEM (CB-SEM and PLS-SEM, respectively). | | | |

The proportions of the races across samples showed that the US OCM and the worldwide OCM sample differed the least. Additionally, the US and worldwide OCM samples differed the least from the student sample. The proportions of races differed most between the non-US OCM sample and all others.

The results for family structure show that the student sample had a significantly larger proportion of singles with no children. A significantly smaller proportion of married students had children than any OCM sample. Further, compared to the OCM samples, the student sample has a significantly smaller proportion of married participants without children. The non-US OCM sample had a significantly larger proportion of singles without children (28.9%) than the US OCM or worldwide OCM.

Significant differences existed in the distributions and means in participants' incomes between all samples. On average, incomes were highest among the US OCM participants, followed by the worldwide OCM participants, the non-US OCM group, and the student participants.

## 3.2 Timing and Attentiveness

Significant differences existed across samples in the average time to complete the survey. Many respondents had sizable differences between the total time spent with the survey open in their browser and the time spent on pages with questions. Therefore, we measured the time participants spent on the relevant survey sections (the videos, the UTAUT 2 items, and the demographic items)—the time from when consent was given to after they answered the last demographic question. We found that the US OCM and student sample groups completed the survey in the least amount of time on average, and the non-US OCM group took the longest. Regressing the natural log of the duration on sample dummies revealed that only the non-US OCM sample differed significantly in mean duration from the other sample groups.

Differences existed between our OCM and student samples regarding attention-check questions. Statistical tests revealed significant variations in the proportions of sample groups that passed the embedded attention check and the price check. Logistic regression analysis showed that the OCM samples did not significantly differ in passing the embedded attention check but that the student participants had statistically significantly lower odds of passing than the OCM samples. Regarding the price-check question, logistic regression indicated no significant differences between the student and non-US OCM samples and between the US and worldwide OCM samples. However, there were significant

[4] The complete set of results is available upon request from the authors.

differences between the two groups. Both students and non-US OCM samples had significantly better odds of passing the price check than the US or worldwide OCM samples.

**Table 4. Sample Demographics**

| | | | | OCM | | |
|---|---|---|---|---|---|---|
| | | TOTAL | Students | US | Non-US | Worldwide |
| | SAMPLE SIZE | 1612 (792) | 131 (165) | 498 (222) | 488 (212) | 495 (193) |
| Gender+ | Male | 0.62 | 0.47 (0.56) | 0.64 (0.43) | 0.69 (0.72) | 0.59 (0.70) |
| | Other | 0.38 | 0.53 (0.42) | 0.36 (0.57) | 0.31 (0.28) | 0.41 (0.30) |
| Age | Mean | 35.58 | 22.18 (23.00) | 39.62 (32.00) | 33.69 (29.00) | 36.93 (29.00) |
| | Median | 33.00 | 20.00 (21.00) | 36.50 (28.00) | 32.00 (27.00) | 34.00 (26.00) |
| | Min. | 19.00 | 19.00 (18.00) | 20.00 (16.00) | 19.00 (17.00) | 19.00 (18.00) |
| | Max. | 85.00 | 54.00 (48.00) | 76.00 (68.00) | 74.00 (63.00) | 85.00 (62.00) |
| Education Level+ | Less Than Highschool | 0.00 | 0.00 (0.00) | 0.01 (0.01) | 0.00 (0.00) | 0.00 (0.01) |
| | Highschool/GED | 0.05 | 0.18 (0.08) | 0.03 (0.11) | 0.04 (0.07) | 0.04 (0.10) |
| | Some College | 0.13 | 0.66 (0.63) | 0.08 (0.35) | 0.06 (0.12) | 0.10 (0.13) |
| | 2year College Degree | 0.06 | 0.11 (0.08) | 0.05 (0.12) | 0.05 (0.07) | 0.07 (0.10) |
| | 4yr College Degree | 0.49 | 0.03 (0.14) | 0.53 (0.32) | 0.53 (0.42) | 0.52 (0.36) |
| | Masters | 0.25 | 0.02 (0.05) | 0.29 (0.07) | 0.27 (0.31) | 0.26 (0.29) |
| | Doctoral | 0.01 | 0.00 (0.00) | 0.00 (0.01) | 0.02 (0.02) | 0.00 (0.01) |
| | Prof (JD, MD) | 0.01 | 0.00 (0.00) | 0.01 (0.01) | 0.02 (0.02) | 0.00 (0.01) |
| Race+ | White/ Caucasian | 0.47 | 0.73 (0.73) | 0.59 (0.69) | 0.22 (0.11) | 0.52 (0.25) |
| | Black | 0.18 | 0.08 (0.04) | 0.29 (0.09) | 0.03 (0.00) | 0.24 (0.01) |
| | Hispanic | 0.04 | 0.04 (0.04) | 0.02 (0.03) | 0.06 (0.02) | 0.05 (0.02) |
| | Asian | 0.26 | 0.05 (0.11) | 0.06 (0.07) | 0.66 (0.83) | 0.13 (0.61) |
| | Indigenous | 0.00 | 0.00 (0.01) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.01) |
| | Pacific Islander | 0.00 | 0.01 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.01) |
| | Prefer To Self-Describe | 0.01 | 0.01 | 0.00 | 0.02 | 0.01 |
| | Pref Not to Say | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 |
| | Multiracial | 0.03 | 0.08 (0.05) | 0.03 (0.10) | 0.01 (0.04) | 0.04 (0.10) |
| Family Structure+ | Single W/Out Children | 0.26 | 0.84 (0.84) | 0.17 (0.50) | 0.29 (0.51) | 0.18 (0.48) |
| | Single W/ Children | 0.03 | 0.01 (0.00) | 0.03 (0.07) | 0.01 (0.01) | 0.04 (0.02) |
| | Married W/Out Children | 0.11 | 0.03 (0.07) | 0.10 (0.14) | 0.11 (0.18) | 0.15 (0.13) |
| | Married W/ Children | 0.51 | 0.07 (0.04) | 0.59 (0.19) | 0.49 (0.26) | 0.55 (0.28) |
| | Life Partner W/Out Children | 0.04 | 0.05 (0.04) | 0.04 (0.05) | 0.05 (0.01) | 0.03 (0.06) |
| | Life Partner W/ Children | 0.05 | 0.00 (0.00) | 0.06 (0.05) | 0.05 (0.02) | 0.04 (0.03) |
| Income+ In USD | < $19,999 | 0.23 | 0.75 (0.73) | 0.07 (0.29) | 0.36 (0.58) | 0.11 (0.52) |
| | $20,000 - $29,999 | 0.14 | 0.05 (0.07) | 0.10 (0.21) | 0.20 (0.18) | 0.15 (0.18) |
| | $30,000 - $39,999 | 0.10 | 0.03 (0.06) | 0.09 (0.17) | 0.11 (0.10) | 0.11 (0.07) |
| | $40,000 - $49,999 | 0.13 | 0.02 (0.01) | 0.18 (0.08) | 0.09 (0.05) | 0.15 (0.10) |
| | $50,000 - $59,999 | 0.15 | 0.02 (0.01) | 0.21 (0.10) | 0.07 (0.01) | 0.19 (0.04) |
| | $60,000 - $69,999 | 0.07 | 0.04 (0.02) | 0.09 (0.03) | 0.05 (0.02) | 0.07 (0.03) |
| | $70,000 - $79,999 | 0.07 | 0.08 (0.01) | 0.08 (0.02) | 0.02 (0.03) | 0.08 (0.02) |
| | $80,000 - $89,999 | 0.07 | 0.02 (0.04) | 0.10 (0.04) | 0.05 (0.02) | 0.08 (0.02) |
| | > = 90,000 | 0.00 | 0.00 (0.01) | 0.00 (0.04) | 0.00 (0.02) | 0.00 (0.00) |
| Time Spent In minutes | Mean | 20.51 | 19.24 | 19.76 | 22.01 | 20.11 |
| | Median | 17.62 | 16.85 | 16.84 | 19.60 | 17.01 |
| | Min. | 6.58 | 7.48 | 6.58 | 7.70 | 6.74 |
| | Max. | 122.53 | 70.24 | 115.96 | 69.04 | 122.53 |

Notes: Where available, Steelman et al.'s (2014) results are shown in parentheses.
+ Values are the proportions of their respective sample.

## 3.3 Comparison of Demographics

A comparison of the current findings and those from Steelman et al. (2014) is shown in Table 6. Unlike the original study, we found that only the student sample was primarily female, reflecting the ongoing demographic shift away from male majorities at US universities (Belkin, 2021; Duffin, 2021; Georgetown University, 2021; Reeves & Smith, 2021). We found few demographic similarities with the work of Steelman et al. (2014). They found their US OCM sample to be the most like their student sample; we found the worldwide OCM and US OCM samples were equally like the student sample. Steelman et al.

(2014) found their non-US OCM and worldwide OCM samples to be the most alike. Our results indicated that the US OCM and worldwide OCM samples were more alike than the non-US OCM was with either the worldwide OCM or the US OCM samples. Finally, the original study found that their worldwide OCM and non-US OCM participants had higher levels of education; we found that participants in the US and non-US OCM samples were the most educated on average.

**Table 5. Demographic Differences**

| | | US OCM vs. Students | Non-US OCM vs. Students | Worldwide OCM vs. Students | US OCM vs. Non-US OCM | US OCM vs. Worldwide OCM | Worldwide OCM vs. Non-US OCM |
|---|---|---|---|---|---|---|---|
| Gender[1] | Male | ***12.15****** | ***20.84****** | ***5.33**** | 2.54 | 3.13 | ***11.20****** |
| | Other | | | | | | |
| Age[2] | Mean Difference | ***17.44****** | ***11.52****** | ***14.75****** | ***5.92****** | **2.67***** | ***3.23****** |
| Education Level | Mean Difference[2] | ***1.98****** | ***2.00****** | ***1.82****** | 0.02 | **0.16*** | ***-0.17**** |
| | Categorical Rank[3] | ***-15.19****** | ***-14.89****** | ***-14.27****** | 0.34 | **-2.13*** | ***2.41**** |
| Race[1] | White/Caucasian | ***8.94****** | ***121.02****** | ***18.54****** | ***137.37****** | **4.53*** | ***94.65****** |
| | Black | ***23.87****** | ***9.11**** | ***15.10****** | ***128.17****** | 3.55 | ***95.39****** |
| | Hispanic | 2.50 | 0.62 | 0.17 | ***11.10**** | **7.59**** | 0.40 |
| | Asian | 0.22 | ***154.12****** | ***7.80**** | ***387.90****** | **17.22***** | ***281.21****** |
| | Indigenous | 0.26 | NA | 0.53 | 0.98 | 0.34 | 1.98 |
| | Pacific Islander | 3.81 | 3.73 | 3.79 | NA | NA | NA |
| | Prefer To Self-Describe | 0.29 | 0.55 | 0.04 | 3.76 | 0.21 | 2.97 |
| | Pref Not to Say | 1.03 | 0.00 | 3.78 | 1.87 | 1.00 | ***4.07**** |
| | Multiracial | ***4.50**** | ***19.08****** | 2.17 | ***6.45**** | 0.70 | ***10.76**** |
| | Invalid Response | 0.53 | 0.54 | 0.53 | 0.00 | 0.00 | 0.00 |
| Family Structure[1] | Single W/out Children | ***215.11****** | ***129.95****** | ***206.21****** | ***18.79****** | 0.142 | ***15.68****** |
| | Single W/ Children | 3.62 | 0.073 | ***4.00**** | ***6.45**** | 0.70 | ***10.76**** |
| | Married W/out Children | ***6.98**** | ***7.81**** | ***12.83****** | 0.10 | **3.83*** | 2.66 |
| | Married W/ Children | ***113.89****** | ***77.53****** | ***96.79****** | ***9.64**** | 1.86 | 3.05 |
| | Life Partner W/out Children | 0.61 | 0.16 | 1.02 | 0.30 | 0.10 | 0.74 |
| | Life Partner W/ Children | ***8.00**** | ***6.99**** | ***6.03**** | 0.23 | 0.97 | 0.25 |
| Income | Mean Difference[2] | ***2.73****** | ***0.84****** | ***2.29****** | ***1.89****** | 0.44** | ***1.44****** |
| | Categorical Rank[3] | ***-11.63****** | ***-6.18****** | ***-10.65****** | ***-12.87****** | -3.12** | ***-10.16****** |
| Time Spent[2] | Mean Difference[2] | 0.52 | ***2.77**** | .87 | ***-2.25**** | -0.36 | ***-1.89**** |

Notes: Sig: * p < 0.05; ** p < 0.01, *** p < 0.001
[1]Chi-square proportion.
[2] Mean difference t-test w/Welch's unequal variance correction.
[3] Wilcoxon rank-sum test Z-stat

The answer to some of these findings may be found in the breakdown of the samples by country. Steelman et al. (2014) found that their worldwide OCM sample primarily comprised individuals from India; however, around 84% of our worldwide OCM respondents were from the United States. The prevalence of US participants in the worldwide OCM sample provides evidence to the cause of many, if not all, of the similarities between these samples (and the lack thereof found with Steelman et al. (2014)). The Mechanical Turk platform is ever evolving. Therefore, researchers should keep current with the existing demographic breakdowns[5] (Difallah et al., 2018). Additionally, the speed of data collection on OCMs makes it vital to keep the time of day (and the time zone of your target population) in mind when releasing a survey, as it could affect which individuals in the global population are more likely to see the study.

## 3.4 Common Method Variance Assessment

We conducted two tests to check for common method variance (CMV or CMB) in our data. First, the Harman single-factor test and Kock's single-factor test for PLS-SEM (Kock, 2020; Podsakoff et al., 2003) indicated no CMV issues. The second test was the stricter full collinearity variance inflation factor test

[5] https://demographics.mturk-tracker.com/

(FCVIF) (Kock, 2015). We found none of the samples exceeded the VIF threshold of 3.3, providing further confidence in the absence of CMV impacting our results.

**Table 6. Comparison of Demographics Findings**

| Steelman et al. 2014 Findings | Current study's findings |
|---|---|
| *"In terms of gender, the … US OCM sample consisted of more women (… 57.2%…)*" (p. 361) | Only the student sample was majority female. |
| *"…the worldwide OCM, non-US OCM, and student samples all consisted of more men (70.5%, 72.1%, and 56%, respectively)."* (p.361) | |
| *"The education levels of participants in the worldwide and non-US OCM samples tended to be higher than all other samples while those of the … US OCM sample were not significantly different."* (p. 361) | All samples have significantly different means and distributions of education except for the US OCM and non-US OCM samples, both of which are the most educated samples on average. The non-US OCM sample has the largest proportions in the top educational categories. |
| *"… interestingly, we found the student, … and US OCM samples were highly similar in many of the demographic distributions…* (p.361) | The student sample differed from the OCM samples in all demographic measures. While the student sample had the most demographic similarities with the US and worldwide OCM samples, there were more non-significant demographic differences between these samples than in Steelman et al. (2014). |
| *"The demographic distributions of the worldwide and non-US OCM samples are highly similar …"* (p.361) | The US and worldwide OCM samples have the most demographic similarities. |
| *"Additionally, the demographic data for the students … and US OCM samples were all fairly similar across many of the categories with the primary differences being the age, education levels, income, and family compositions."* (p.362) | Of the OCM samples, the worldwide and US samples were the most like the student sample. The only difference among these differences is that the US OCM sample had more Black participants than the student sample, and the worldwide OCM sample had more married, childless participants than the student sample. |

## 3.5 Structural Model Analyses

We used partial least squares structural equation modeling (PLS-SEM)[6] to analyze the UTAUT 2 model. We used PLS-SEM for analysis because of the complexity of the UTAUT 2 model and the number of relationships within the model (Fornell & Bookstein, 1982; Hair Jr., Black, et al., 2019). Additionally, the student sample size of 131 was below the recommended 20 observations per indicator for covariance-based SEM (Hair Jr. et al., 2019) but above the minimum of five per variable in the largest regression in the model for PLS-SEM (Hair Jr. et al., 2022). Finally, UTAUT and UTAUT 2 were developed and verified using PLS-SEM (Venkatesh et al., 2003, 2012). Because UTAUT 2 contains many moderators—and those moderation effects are not consistently significant in follow-up research (e.g., Tamilmani et al., 2021; Venkatesh et al., 2016)—we tested and reported only direct effects that have remained stable and context-free compared to the interaction effects. Because actual usage was not accurately measured, due to the newness of the technology at the time, and the goal was to compare samples, not to support or refute UTAUT 2, we dropped the use construct from our model.

### 3.5.1 Factor Analysis

Table 7 presents each sample's reliability estimates, and average variance extracted (AVE). Table 8 presents the correlation matrices. The student sample had adequate reliability, convergent validity, and discriminant validity. All reliability estimates for the student sample were above the recommended threshold ( > 0.70); the AVEs, as well ( > 0.50), and the square root of the AVEs all exceeded the correlations with other constructs (Fornell & Larcker, 1981).

The US OCM sample had adequate reliability and convergent validity, but no discriminant validity between effort expectancy and facilitating conditions. We note that there were high correlations between effort

[6] SmartPLS 3 and 4 were used for the analysis, and all significance statistics were based on 5,000 bootstrapped samples (Ringle et al., 2015, 2022).

expectancy and facilitating conditions in all samples and that this correlation exceeded the explained variance only in the US OCM sample.

For the non-US OCM sample, all reliability estimates were above the recommended thresholds except for the reliability and consistency measures of the facilitating conditions construct. Only the composite reliability score exceeded the 0.70 threshold. Additionally, despite the AVE for facilitating conditions being below the 0.50 threshold, its square root was larger than any off-diagonal correlation. This provided evidence of discriminant validity despite the low convergent validity and low internal consistency reliability.

Like the student sample, the worldwide OCM sample had acceptable reliability and validity measures. All reliability and consistency estimates were above the recommended thresholds. The square roots of the AVEs were larger than any off-diagonal correlation, supporting sufficient convergent and discriminant validity.

### 3.5.2 Heterotrait-Monotrait Ratios

Table 9 shows the heterotrait-monotrait (HTMT)[7] ratios (Henseler et al., 2015) for all the samples. There were no discriminant validity issues with the student or the non-US OCM samples (no values above 0.9). However, the US and worldwide OCM samples had similar patterns of a lack of discrimination between facilitating conditions and effort expectancy. Additionally, the US OCM sample showed a lack of discrimination between habit and behavioral intention. Further, the US OCM and worldwide OCM samples had discriminant validity issues with facilitating conditions and effort expectancy. The non-US OCM sample had reliability and consistency issues with facilitating conditions. Following the minimal cleaning philosophy of Steelman et al. (2014), we did not engage in further efforts (e.g., removing outliers or bad indicators) to improve the psychometrics.

### 3.5.3 Path Coefficients

After assessing the reliability and validity of the model, we estimated the structural model. Path coefficients are reported in Table 10. Examining each construct across samples, we noted several similarities and differences. None of the controls (age, gender, or experience) significantly impacted behavioral intention. In contrast, habit and hedonic motivation were significant ($p < 0.001$) positive predictors of behavioral intention for all samples. Effort expectancy had no statistically significant impact on behavioral intention. Facilitating conditions were significant on behavioral intention only for the non-US OCM sample. Price value was only a significant driver of behavioral intention for the worldwide OCM sample. At the same time, social influence only proved to be a driver for the behavioral intention in the non-US OCM sample. Lastly, performance expectancy was a significant positive determinant of behavioral intention for all but the student sample. Differences in the patterns of statistically significant coefficients involved the non-US OCM (performance expectancy and facilitating conditions), worldwide OCM (price value), and student (performance expectancy) samples.

To test differences between the path coefficients in each model, we used a two-tailed *t*-test of differences (Chin, 2002), as shown in Table 11. Most structural paths had no significant differences across samples, indicating equality of the theoretical relationships between constructs. All the differences were between the non-US OCM and other OCM samples. While the US OCM, students, and worldwide OCM samples did not differ significantly, indicating a potential cultural component—the worldwide OCM and US OCM samples had primarily US participants and were demographically similar—that is not captured by the UTAUT 2 model. We also assessed the predictive power of the different samples. The adjusted $R^2$ values varied from ~.64 to ~.75, but we found no statistically significant pairwise differences between the samples.

[7] Hair et al. (2018, 2022) recommend using the HTMT ratio to assess the discriminant validity of PLS-SEM constructs (i.e., composites). While, Fornell-Larcker's sensitivity decreases for sample sizes over 100 and loadings between 0.6 and 0.8, the HTMT ratio has over 95% sensitivity regardless of sample size and range of loadings (Henseler et al., 2015). Hair et al. (2022) recommend an HTMT ratio below 0.9 for similar empirical measures and 0.85 for different ones.

**Table 7. Reliability and Convergent Validity**

| Students | μ | σ | α | ρA | ρC | AVE |
|---|---|---|---|---|---|---|
| Behavioral Intention | 4.65 | 1.52 | 0.90 | 0.90 | 0.94 | 0.83 |
| Effort Expectancy | 5.4 | 1.13 | 0.91 | 0.95 | 0.94 | 0.78 |
| Facilitating Conditions | 5.34 | 1.04 | 0.76 | 0.80 | 0.84 | 0.57 |
| Habit | 3.82 | 1.40 | 0.86 | 0.88 | 0.91 | 0.71 |
| Hedonic Motivation | 5.87 | 1.10 | 0.95 | 0.95 | 0.97 | 0.91 |
| Performance Expectancy | 3.97 | 1.35 | 0.90 | 0.92 | 0.93 | 0.77 |
| Price Value | 4.92 | 1.09 | 0.89 | 0.93 | 0.93 | 0.82 |
| Social Influence | 3.57 | 1.48 | 0.95 | 0.95 | 0.97 | 0.90 |
| Age | 22.18 | 6.23 | 1.00 | 1.00 | 1.00 | 1.00 |
| Experience | 1.82 | 0.38 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gender | 0.47 | 0.50 | 1.00 | 1.00 | 1.00 | 1.00 |
| **US OCM** | **μ** | **σ** | **α** | **ρA** | **ρC** | **AVE** |
| Behavioral Intention | 5.68 | 1.04 | 0.79 | 0.79 | 0.88 | 0.71 |
| Effort Expectancy | 5.73 | 0.89 | 0.82 | 0.83 | 0.88 | 0.66 |
| Facilitating Conditions | 5.67 | 0.84 | 0.72 | 0.74 | 0.83 | 0.55 |
| Habit | 5.35 | 1.13 | 0.84 | 0.85 | 0.89 | 0.68 |
| Hedonic Motivation | 5.95 | 0.89 | 0.81 | 0.81 | 0.89 | 0.72 |
| Performance Expectancy | 5.41 | 1.21 | 0.88 | 0.88 | 0.92 | 0.74 |
| Price Value | 5.62 | 1.05 | 0.85 | 0.85 | 0.91 | 0.76 |
| Social Influence | 5.23 | 1.37 | 0.91 | 0.91 | 0.94 | 0.84 |
| Age | 39.62 | 11.18 | 1.00 | 1.00 | 1.00 | 1.00 |
| Experience | 1.34 | 0.47 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gender | 0.64 | 0.48 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Non-US OCM** | **μ** | **σ** | **α** | **ρA** | **ρC** | **AVE** |
| Behavioral Intention | 5.77 | 0.93 | 0.81 | 0.81 | 0.89 | 0.72 |
| Effort Expectancy | 5.97 | 0.76 | 0.83 | 0.83 | 0.88 | 0.66 |
| Facilitating Conditions | 5.69 | 0.80 | ***0.66*** | ***0.67*** | 0.79 | ***0.49*** |
| Habit | 5.3 | 1.10 | 0.86 | 0.87 | 0.91 | 0.71 |
| Hedonic Motivation | 6.27 | 0.79 | 0.84 | 0.84 | 0.90 | 0.76 |
| Performance Expectancy | 5.6 | 1.13 | 0.91 | 0.91 | 0.93 | 0.78 |
| Price Value | 5.58 | 1.12 | 0.89 | 0.91 | 0.93 | 0.82 |
| Social Influence | 5.24 | 1.34 | 0.92 | 0.92 | 0.95 | 0.86 |
| Age | 33.69 | 7.93 | 1.00 | 1.00 | 1.00 | 1.00 |
| Experience | 1.59 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gender | 0.69 | 0.46 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Worldwide OCM** | **μ** | **σ** | **α** | **ρA** | **ρC** | **AVE** |
| Behavioral Intention | 5.76 | 0.91 | 0.75 | 0.75 | 0.86 | 0.67 |
| Effort Expectancy | 5.74 | 0.90 | 0.83 | 0.83 | 0.89 | 0.66 |
| Facilitating Conditions | 5.72 | 0.87 | 0.76 | 0.78 | 0.85 | 0.58 |
| Habit | 5.39 | 1.06 | 0.85 | 0.86 | 0.90 | 0.69 |
| Hedonic Motivation | 6.04 | 0.87 | 0.83 | 0.83 | 0.90 | 0.74 |
| Performance Expectancy | 5.57 | 1.05 | 0.86 | 0.87 | 0.90 | 0.70 |
| Price Value | 5.65 | 0.99 | 0.83 | 0.83 | 0.90 | 0.75 |
| Social Influence | 5.27 | 1.25 | 0.90 | 0.90 | 0.94 | 0.83 |
| Age | 36.93 | 10.98 | 1.00 | 1.00 | 1.00 | 1.00 |
| Experience | 1.35 | 0.48 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gender | 0.59 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 |

μ : Mean
σ : Standard Deviation
α : Cronbach's alpha
ρA : Reliability Coefficient
ρC : Composite Reliability
AVE : Average Variance Extracted

**Table 8. Construct Correlations**

| | **Students** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Behavioral Intention | 0.91 | | | | | | | | | |
| 2 | Effort Expectancy | 0.34 | 0.89 | | | | | | | | |
| 3 | Facilitating Conditions | 0.36 | 0.60 | 0.76 | | | | | | | |
| 4 | Habit | 0.71 | 0.25 | 0.21 | 0.84 | | | | | | |
| 5 | Hedonic Motivation | 0.52 | 0.63 | 0.51 | 0.26 | 0.95 | | | | | |
| 6 | Performance Expectancy | 0.63 | 0.23 | 0.28 | 0.62 | 0.43 | 0.88 | | | | |
| 7 | Price Value | 0.37 | 0.37 | 0.29 | 0.41 | 0.36 | 0.38 | 0.91 | | | |
| 8 | Social Influence | 0.63 | 0.27 | 0.24 | 0.68 | 0.28 | 0.60 | 0.34 | 0.95 | | |
| 9 | Age | -0.02 | -0.05 | -0.14 | -0.05 | -0.02 | 0.07 | 0.21 | -0.10 | NA | |
| 10 | Experience | 0.02 | -0.25 | -0.18 | 0.15 | -0.20 | 0.06 | 0.05 | 0.07 | 0.00 | NA |
| | Gender | -0.05 | 0.13 | 0.14 | -0.04 | -0.04 | -0.14 | -0.12 | -0.01 | -0.07 | -0.17 |
| | **US OCM** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | Behavioral Intention | 0.84 | | | | | | | | | |
| 2 | Effort Expectancy | 0.57 | 0.81 | | | | | | | | |
| 3 | Facilitating Conditions | 0.58 | ***0.76*** | ***0.74*** | | | | | | | |
| 4 | Habit | 0.78 | 0.56 | 0.56 | 0.82 | | | | | | |
| 5 | Hedonic Motivation | 0.68 | 0.55 | 0.57 | 0.52 | 0.85 | | | | | |
| 6 | Performance Expectancy | 0.72 | 0.45 | 0.46 | 0.75 | 0.49 | 0.86 | | | | |
| 7 | Price Value | 0.61 | 0.57 | 0.56 | 0.64 | 0.51 | 0.58 | 0.87 | | | |
| 8 | Social Influence | 0.61 | 0.40 | 0.44 | 0.73 | 0.36 | 0.77 | 0.54 | 0.92 | | |
| 9 | Age | -0.07 | -0.15 | -0.18 | -0.07 | -0.07 | -0.02 | -0.08 | -0.05 | NA | |
| 10 | Experience | -0.19 | -0.13 | -0.16 | -0.35 | 0.06 | -0.34 | -0.32 | -0.45 | 0.12 | NA |
| | Gender | 0.04 | 0.12 | 0.12 | 0.14 | 0.02 | 0.06 | 0.11 | 0.15 | -0.10 | -0.20 |
| | **Non-US OCM** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | Behavioral Intention | 0.85 | | | | | | | | | |
| 2 | Effort Expectancy | 0.53 | 0.81 | | | | | | | | |
| 3 | Facilitating Conditions | 0.54 | 0.62 | 0.70 | | | | | | | |
| 4 | Habit | 0.71 | 0.34 | 0.41 | 0.84 | | | | | | |
| 5 | Hedonic Motivation | 0.66 | 0.51 | 0.41 | 0.43 | 0.87 | | | | | |
| 6 | Performance Expectancy | 0.70 | 0.42 | 0.45 | 0.68 | 0.46 | 0.88 | | | | |
| 7 | Price Value | 0.51 | 0.37 | 0.45 | 0.55 | 0.35 | 0.59 | 0.90 | | | |
| 8 | Social Influence | 0.63 | 0.38 | 0.41 | 0.65 | 0.30 | 0.80 | 0.57 | 0.93 | | |
| 9 | Age | -0.04 | -0.01 | -0.01 | -0.08 | 0.03 | 0.03 | 0.05 | -0.01 | NA | |
| 10 | Experience | -0.08 | 0.03 | -0.12 | -0.29 | 0.22 | -0.18 | -0.18 | -0.33 | 0.18 | NA |
| | Gender | 0.04 | 0.02 | 0.05 | 0.01 | 0.02 | 0.00 | -0.01 | -0.03 | -0.01 | -0.03 |
| | **Worldwide OCM** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | Behavioral Intention | 0.82 | | | | | | | | | |
| 2 | Effort Expectancy | 0.61 | 0.81 | | | | | | | | |
| 3 | Facilitating Conditions | 0.59 | 0.74 | 0.76 | | | | | | | |
| 4 | Habit | 0.72 | 0.52 | 0.48 | 0.83 | | | | | | |
| 5 | Hedonic Motivation | 0.64 | 0.64 | 0.52 | 0.44 | 0.86 | | | | | |
| 6 | Performance Expectancy | 0.70 | 0.55 | 0.52 | 0.74 | 0.49 | 0.84 | | | | |
| 7 | Price Value | 0.66 | 0.60 | 0.62 | 0.60 | 0.59 | 0.59 | 0.87 | | | |
| 8 | Social Influence | 0.60 | 0.51 | 0.50 | 0.73 | 0.35 | 0.72 | 0.61 | 0.91 | | |
| 9 | Age | -0.06 | -0.08 | -0.07 | -0.13 | 0.00 | -0.03 | -0.06 | -0.11 | NA | |
| 10 | Experience | -0.13 | -0.09 | -0.13 | -0.30 | 0.18 | -0.23 | -0.17 | -0.38 | 0.12 | NA |
| | Gender | 0.05 | 0.08 | 0.04 | 0.11 | -0.01 | 0.05 | 0.06 | 0.11 | -0.10 | -0.21 |
| Note: The square root of the AVE appears on the diagonals. | | | | | | | | | | | |

**Table 9. Heterotrait-Monotrait (HTMT) Ratios**

| **Students** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AGE | BI | EE | EXP | FC | GEN | HAB | HM | PE | PV |
| Behavioral Intention | 0.063 | | | | | | | | | |
| Effort Expectancy | 0.054 | 0.365 | | | | | | | | |
| Experience | 0.003 | 0.068 | 0.261 | | | | | | | |
| Facilitating Conditions | 0.174 | 0.414 | 0.700 | 0.227 | | | | | | |
| Gender | 0.073 | 0.076 | 0.142 | 0.165 | 0.163 | | | | | |
| Habit | 0.085 | 0.787 | 0.274 | 0.162 | 0.243 | 0.058 | | | | |
| Hedonic Motivation | 0.022 | 0.570 | 0.670 | 0.206 | 0.591 | 0.044 | 0.279 | | | |
| Performance Expectancy | 0.070 | 0.678 | 0.234 | 0.093 | 0.320 | 0.165 | 0.692 | 0.450 | | |
| Price Value | 0.222 | 0.403 | 0.407 | 0.048 | 0.342 | 0.131 | 0.460 | 0.379 | 0.410 | |
| Social Influence | 0.102 | 0.676 | 0.291 | 0.071 | 0.278 | 0.022 | 0.754 | 0.297 | 0.647 | 0.362 |
| **US OCM** | | | | | | | | | | |
| | AGE | BI | EE | EXP | FC | GEN | HAB | HM | PE | PV |
| Behavioral Intention | 0.077 | | | | | | | | | |
| Effort Expectancy | 0.169 | 0.696 | | | | | | | | |
| Experience | 0.116 | 0.218 | 0.145 | | | | | | | |
| Facilitating Conditions | 0.207 | 0.752 | ***0.981*** | 0.195 | | | | | | |
| Gender | 0.096 | 0.045 | 0.128 | 0.197 | 0.137 | | | | | |
| Habit | 0.074 | ***0.957*** | 0.660 | 0.383 | 0.697 | 0.150 | | | | |
| Hedonic Motivation | 0.079 | 0.848 | 0.675 | 0.085 | 0.737 | 0.018 | 0.621 | | | |
| Performance Expectancy | 0.021 | 0.857 | 0.522 | 0.365 | 0.559 | 0.070 | 0.865 | 0.575 | | |
| Price Value | 0.085 | 0.745 | 0.678 | 0.344 | 0.707 | 0.124 | 0.750 | 0.619 | 0.665 | |
| Social Influence | 0.048 | 0.720 | 0.455 | 0.473 | 0.537 | 0.160 | 0.834 | 0.416 | 0.859 | 0.613 |
| **Non-US OCM** | | | | | | | | | | |
| | AGE | BI | EE | EXP | FC | GEN | HAB | HM | PE | PV |
| Behavioral Intention | 0.058 | | | | | | | | | |
| Effort Expectancy | 0.036 | 0.642 | | | | | | | | |
| Experience | 0.179 | 0.150 | 0.072 | | | | | | | |
| Facilitating Conditions | 0.071 | 0.719 | 0.848 | 0.150 | | | | | | |
| Gender | 0.010 | 0.047 | 0.047 | 0.032 | 0.066 | | | | | |
| Habit | 0.085 | 0.848 | 0.394 | 0.316 | 0.530 | 0.016 | | | | |
| Hedonic Motivation | 0.035 | 0.805 | 0.616 | 0.237 | 0.551 | 0.026 | 0.505 | | | |
| Performance Expectancy | 0.029 | 0.812 | 0.469 | 0.198 | 0.564 | 0.023 | 0.758 | 0.517 | | |
| Price Value | 0.048 | 0.597 | 0.430 | 0.196 | 0.577 | 0.021 | 0.623 | 0.401 | 0.649 | |
| Social Influence | 0.022 | 0.724 | 0.431 | 0.348 | 0.507 | 0.027 | 0.722 | 0.343 | 0.875 | 0.620 |
| **Worldwide OCM** | | | | | | | | | | |
| | AGE | BI | EE | EXP | FC | GEN | HAB | HM | PE | PV |
| Behavioral Intention | 0.067 | | | | | | | | | |
| Effort Expectancy | 0.084 | 0.776 | | | | | | | | |
| Experience | 0.125 | 0.231 | 0.097 | | | | | | | |
| Facilitating Conditions | 0.078 | 0.766 | ***0.921*** | 0.153 | | | | | | |
| Gender | 0.096 | 0.056 | 0.089 | 0.210 | 0.074 | | | | | |
| Habit | 0.143 | 0.893 | 0.607 | 0.336 | 0.588 | 0.122 | | | | |
| Hedonic Motivation | 0.007 | 0.808 | 0.775 | 0.193 | 0.652 | 0.021 | 0.511 | | | |
| Performance Expectancy | 0.035 | 0.855 | 0.635 | 0.261 | 0.638 | 0.057 | 0.863 | 0.571 | | |
| Price Value | 0.070 | 0.839 | 0.720 | 0.184 | 0.767 | 0.064 | 0.711 | 0.719 | 0.700 | |
| Social Influence | 0.119 | 0.728 | 0.596 | 0.405 | 0.600 | 0.122 | 0.837 | 0.401 | 0.825 | 0.710 |

**Table 10. Standardized Path Coefficients by Sample**

| | Non-US OCM | STUDENT | US OCM | Worldwide OCM |
|---|---|---|---|---|
| AGE → BI | -0.03(1.248) | 0.04(0.785) | -0.01(0.264) | 0.01(0.302) |
| GEN → BI | 0.03(1.332) | 0.00(0.043) | -0.04(1.779) | -0.01(0.414) |
| EXP → BI | 0.04(1.320) | 0.00(0.083) | 0.03(0.723) | 0.00(0.131) |
| FC → BI | ***0.11(2.668)*** ** | 0.10(1.399) | 0.05(0.864) | 0.11(1.639) |
| EE → BI | 0.07(1.615) | -0.10(1.286) | 0.02(0.380) | 0.03(0.545) |
| HAB → BI | ***0.32(8.847)*** *** | ***0.46(4.917)*** *** | ***0.42(7.370)*** *** | ***0.34(5.732)*** *** |
| HM → BI | ***0.32(8.146)*** *** | ***0.34(4.147)*** *** | ***0.30(6.808)*** *** | ***0.25(3.841)*** *** |
| PE → BI | ***0.16(2.857)*** ** | 0.10(0.886) | ***0.22(3.336)*** *** | ***0.18(2.859)*** ** |
| PV → BI | -0.02(0.512) | -0.03(0.542) | 0.05(0.828) | ***0.12(2.476)*** * |
| SI → BI | ***0.15(3.058)*** ** | 0.17(1.449) | 0.00(0.058) | -0.01(0.245) |

Coefficient (t-value) Sig: * - 0.05; ** - 0.01, ***- 0.001 derived from 5000 bootstrapped samples

**Table 11. Path Coefficient Differences**

| | Non-US OCM - Students | Non-US OCM - US OCM | Non-US OCM - Worldwide OCM | Students - US OCM | Students - Worldwide OCM | US OCM - Worldwide OCM |
|---|---|---|---|---|---|---|
| AGE → BI | -0.071(1.281) | -0.024(0.649) | -0.040(1.079) | 0.047(0.822) | 0.031(0.549) | -0.016(0.402) |
| GEN → BI | 0.032(0.539) | ***0.073(2.255)**** | 0.041(1.181) | 0.041(0.675) | 0.009(0.141) | -0.032(0.897) |
| EXP → BI | 0.046(0.656) | 0.015(0.326) | 0.045(1.037) | -0.031(0.427) | -0.001(0.018) | 0.030(0.637) |
| FC → BI | 0.009(0.119) | 0.057(0.876) | 0.006(0.093) | 0.048(0.557) | -0.003(0.031) | -0.051(0.644) |
| EE → BI | 0.170(1.912) | 0.049(0.767) | 0.033(0.442) | -0.121(1.315) | -0.137(1.374) | -0.016(0.205) |
| HAB → BI | -0.133(1.316) | -0.097(1.414) | -0.026(0.363) | 0.036(0.329) | 0.107(0.952) | 0.071(0.840) |
| HM → BI | -0.017(0.187) | 0.027(0.451) | 0.071(0.928) | 0.044(0.471) | 0.088(0.841) | 0.044(0.566) |
| PE → BI | 0.073(0.609) | -0.054(0.623) | -0.013(0.148) | -0.126(1.036) | -0.085(0.705) | 0.041(0.467) |
| PV → BI | 0.018(0.223) | -0.064(0.973) | ***-0.147(2.282)**** | -0.082(0.930) | -0.166(1.892) | -0.083(1.145) |
| SI → BI | -0.042(0.306) | ***0.150(2.164)**** | ***0.160(2.261)**** | 0.193(1.385) | 0.203(1.449) | 0.010(0.137) |

Coefficient difference (t-value) Sig: * - 0.05; ** - 0.01, ***- 0.001 derived from 5000 bootstrapped samples

### 3.5.4 Measurement Invariance Tests

We also tested measurement invariance, as in Steelman et al. (2014). PLS-SEM estimates models using composite constructs. We used the three-step measurement invariance of composite models[8] (MICOM) procedure (Henseler et al., 2016). The first step assesses configural invariance—"Identical indicators per measurement model…Identical data treatment…Identical algorithm settings or optimization criteria…" (Henseler et al., 2016, p. 413). As the data for our samples was collected using an identical survey and analyzed using identical methods, our samples had configural invariance. In the second step, the consistency of the indicators in creating their focal constructs is compared across different samples. The third step examines whether there are any significant differences in the mean and variance of the construct scores between observations from different samples. It is important to note that each step is a prerequisite for the next. Compositional invariance is required for meaningful comparisons of path coefficients and latent variable loadings. Strict measurement invariance—mean and variance—is necessary to combine samples (Henseler et al., 2016).

Table 12 shows a summary of the MICOM test results. These tests indicated compositional measurement invariance between all samples except for the student and non-US OCM samples, which lacked invariance in the composition of the habit construct. Thus, we concluded that multigroup comparisons of construct measurements were statistically meaningful, except for comparisons of habit between the student and non-US OCM samples.

Only the US OCM and worldwide OCM samples came close to achieving step 3 invariance for any construct; they differed in means and variances only for the performance expectancy construct. All other

[8] While those conducting CB-SEM invariance testing of causal models may be more familiar with terms configural, metric, scalar, and complete invariance, recent advances in PLS-SEM (e.g., Henseler et al. 2016) utilize the terms configural, compositional, and composite invariance due to the uniqueness of composite-based models examined in PLS-SEM. For complete details of this approach, we recommend examining Henseler et al. (2016) for the procedure.

sample comparisons had significant differences in multiple constructs. Steelman et al. (2014) also found invariance between the factor loadings and intercepts, equivalent to step 2 of the MICOM procedure, across all samples.

**Table 12. MICOM Testing Results**

| Sample Comparison | Configural | Compositional | Mean | Variance |
|---|---|---|---|---|
| | Step 1 | Step 2 | Step 3a | Step 3b |
| Non-US OCM vs. Students | YES | NO* | NO | NO |
| Non-US OCM vs. US OCM | YES | YES | NO | NO |
| Non-US OCM vs. Worldwide OCM | YES | YES | NO | NO |
| Students vs. US OCM | YES | YES | NO | NO |
| Students vs. Worldwide OCM | YES | YES | NO | NO |
| US OCM vs. Worldwide OCM | YES | YES | NO | NO |
| * Only Habit was not invariant – all other constructs were. | | | | |

## 3.6 Summary

A summary of the results of our analysis and Steelman et al. (2014) appears in Table 13. The non-US OCM differed greatly from the US and worldwide OCM, while the latter were more like the student sample.

**Table 13. Comparison of Analysis Results**

| Criterion | Steelman et al. (2014) Results | Our Results |
|---|---|---|
| Validity & Reliability | All samples' constructs had ρC > 0.85 | All samples exceed thresholds except the non-US OCM sample's Cronbach's α and ρC for facilitating conditions. |
| | All samples' constructs AVEs > 0.50 | All samples exceed thresholds except the non-US OCM sample AVE for facilitating conditions. |
| | All samples' sq. root of AVE > all correlations | All samples exceed thresholds except for facilitating conditions and effort expectancy in the US OCM sample; HTMT analysis showed discriminant validity concerns with all samples except the student sample |
| | All sample items load more on their focal constructs than others | All samples exceed the recommended thresholds |
| Path Coefficient Differences | Significant coefficients for all samples | No sample resulted in all structural paths being significant; the worldwide OCM and non-US OCM samples had 4/10 significant paths, the US OCM sample had 3/10, and the student sample had 2/10 |
| | 5/18 path coefficients across all samples had significant differences (28%) | 4/48 path coefficients across all samples had significant differences (8%) |
| | The student sample differs from the worldwide OCM and non-US OCM samples; the US OCM and worldwide OCM samples differ. | Most path coefficient differences were between the non-US OCM sample and other OCM samples, 2-US OCM, and 2-worldwide OCM |
| Invariance Tests | Behavioral intention differed between US OCM and worldwide OCM samples (based on the mean variable score) | MICOM testing revealed compositional differences between the non-US OCM and student samples for habit; no mean or variance invariance between any samples; the US OCM and worldwide OCM samples differed only in the means and variances of performance expectancy |

# 4 Discussion and Implications

This study compared results from a previous study by Steelman et al. (2014) to see if Mechanical Turk remains a practical substitute for student samples. The findings indicate that OCM samples do not differ widely from student samples regarding validity and reliability but do widely differ regarding demographic heterogeneity. Therefore, OCM samples continue to offer researchers a cost-efficient, reliable data source capable of complementing or substituting for student samples. Table 14 summarizes our findings compared with Steelman et al. (2014).

**Table 14. Comparison of Findings**

| Steelman et al. 2014 Findings | Our Findings |
|---|---|
| Demographic Differences Across Samples | |
| *"…we found strong demographic differences in our worldwide and non-US OCM samples compared with the student, consumer panel, and US OCM samples."* (p. 371) | The OCM samples are primarily male, while the student sample is primarily female. All samples are significantly different in age (US OCM>worldwide OCM>non-US OCM>Student). Only the US OCM and non-US OCM samples are similar in education (US OCM=non-US OCM>worldwide OCM>Student). Income (US OCM > worldwide OCM>non-US OCM>Students, Race (Student=worldwide OCM, US OCM least like any other sample). |
| SEM Model Differences Across Samples | |
| *"Specifically, we found repeated issues with the worldwide and non-US OCM samples in regard to their CFA model fit indices, structural model fit indices, measurement and scale invariances, and, most importantly, the complete lack of significance of the perceived usefulness–behavioral intention relationship in the TAM model"* (p. 370). | The non-US OCM sample had several internal consistency and reliability issues for FC and statistically significant path coefficient differences between the other OCM samples. HTMT analysis indicated discriminant validity issues between FC and EE in the US OCM and worldwide OCM samples. |
| *"…the responses provided by non-US OCM participants clearly provide different conclusions than those of the US populations collected in this study."* (p. 371) | The samples all followed a similar pattern of effects and prediction within the model while differing in the reported levels of key constructs. The path coefficients from the non-US OCM sample differed significantly from those from other OCM samples. The path coefficients derived from the student sample did not differ significantly from any OCM sample. |
| Psychometric Variety Across Samples | |
| *"… the details of these differences beyond demographics and psychometrics, these results might not have been identified as the psychometrics (reliability, convergent validity, divergent validity, and the factor loadings) did not generally differ among the samples and met all validity thresholds within PLS compared to CB-SEM."* (p. 370) | Although we found that the psychometrics did vary across samples, examining the composition of the samples explained the pattern of results. The most heterogeneous sample, in terms of country, the non-US OCM sample responded the least similarly to all other samples. |
| Student Sample Compared to OCM Samples | |
| *"…we typically noticed consistent results between the student… and US OCM samples, indicating the potential for interchangeability among these samples."* (p, 370) | The most consistent relationship across all samples was between the US OCM and worldwide OCM samples. However, both the US OCM and worldwide OCM samples were consistently more like the student sample than the non-US OCM sample. Primarily due to our worldwide OCM sample having a majority of US OCM participants compared to Steelman et al., whose worldwide OCM sample had a majority of Indian participants. |

## 4.1 Recommendations for Researchers and Reviewers

Existing research offers a variety of evaluations, criticisms, and recommendations for researchers collecting data from OCM samples. We compiled and synthesized much of this advice in Table 15 and Table 16. Further, we add and integrate our reporting recommendations with those of Steelman et al. (2014) in Table 17.

Table 16 shows that the literature agrees on how to handle inattentiveness in surveys. Data quality issues stemming from inattentive responses should be guarded against with novel, objective, unbiased attention checks verifying both that the participant is human and attentive/conscientious (Aguinis et al., 2021; Downs et al., 2010; Jia et al., 2017). Recommendations for assessing attentiveness include using survey questions (i.e., attention-check questions) and tools like CAPTCHA.

**Table 15. Summary of Findings from Prior Research Involving OCM Samples**

| Factor | Findings from Literature |
|---|---|
| Participant attentiveness | Some OCM participants are inattentive; most provide quality responses (Peer et al., 2017). |
| | Mechanical Turk workers are more likely than students to be distracted and rushed when completing HITs (Aguinis et al., 2021). |
| | Some OCM participants are interested in collecting the incentive(s) and do not answer conscientiously (Downs et al., 2010). |
| Validity of OCM samples | When researchers employ the appropriate measures, OCM data is equivalent to (at least as good as) other sampling techniques (Lowry et al., 2016). |
| | US OCM samples are appropriate for academic studies of the US population. (Steelman et al., 2014). |
| | Many OCM participants do not work in a traditional context, so OCM responses may not provide data valid for organizational research (Keith et al., 2019). |
| | "MTurk enhances external validity by targeting all ages, genders, income levels, and educational backgrounds, as well as facilitating participant segmentation (e.g., a specific income bracket, educational cohort, or ethnocultural group)." (Daly & Nataraajan, 2015, p. 2607) |
| Demographics | The demographic distribution of OCM participants is more varied than college students (Steelman et al., 2014). |
| | OCM samples are more diverse than US college student samples (Buhrmester et al., 2011). |
| | OCM samples are valid for "generalized" samples of a broad population. Specific contexts require diligence to ensure sample representativeness. (Jia et al., 2017). |
| | OCM participants may be more representative of non-professional populations for US samples (Buchheit et al., 2018). |
| Financial and social traits | Like students, OCM participants have different social and financial values than the typical population (Goodman et al., 2013). |
| Language & culture | Data collected from non-English-speaking countries have only configural invariance with data from English-speaking countries (Aguinis et al., 2021). |
| | The amount of several types of biases, shirking, and commitment will vary by culture. (Fang et al., 2016). |
| | The impact of the distinct types of problematic responses will vary across samples from distinct cultures (Fang et al., 2016). |
| | Language and culture influence interpretations, affecting results and generalizability (Feitosa et al., 2015). |
| Non-US samples | Because of the considerable variation in cultures and countries, non-US OCM samples may not generalize to a specific population of interest (Steelman et al., 2014). |
| | Results explicitly derived from non-US OCM participants' responses differ from those from US OCM participants (Steelman et al., 2014). |
| Data collection | OCMs provide access to otherwise inaccessible participants (Mason & Suri, 2012). |
| | The population of OCM participants, unlike students, is always readily available (Mason & Suri, 2012). |
| | OCMs provide a more convenient, efficient way to collect samples across locations, cultures, and demographics. And to select and target participants based on these factors (Lowry et al., 2016). |
| | OCM samples can access uncommon participants, allowing tailored samples. (Lowry et al., 2016). |
| | OCMs are less expensive than commercial research panels for the equivalent number of responses (Daly & Nataraajan, 2015). |

There were statistically significant differences in timing and attentiveness. First, the non-US OCM participants had longer average completion times than the other samples, possibly due to varying numbers of native English speakers. The longer survey duration may suggest that non-native or non-US English speakers needed more effort to understand the survey. The non-US sample had the most differences in responses, indicating language may have influenced how participants interpreted the study materials. Second, we found that a significantly lower proportion of student participants passed the embedded attention check question. For participation in the survey, some students were offered extra credit while OCM participants were paid and warned that inattentiveness would result in not being paid. These two incentive structures—gain-only vs. gain and loss—may explain differences in attentiveness between the OCM and student samples. Finally, students and non-US OCM participants passed the price check question at a higher rate than US and worldwide OCM samples, possibly due to their relative youth and lower incomes, making them more likely to be attracted to the technology, price-conscious, or both. Whatever the explanation, MICOM testing found variations in how participants responded to survey items. Had our goal been a UTAUT2 study, controlling for these differences or analyzing each sample in isolation would have been necessary.

**Table 16. Recommendations**

| Factor | Recommendation |
|---|---|
| General | Maintain a list of worker IDs. Use this list to filter bad participants, prevent their involvement in future studies, and filter or manage those who participate multiple times when collecting multiple waves of data within the same study (Aguinis et al., 2021; Jia et al., 2017). |
| | Author or revise study items and scales to use impartial wording to limit biases (Fang et al., 2016; Jia et al., 2017) |
| Attention Checks | Use at least two attention check questions to verify that participants are conscientious, more for longer surveys, and spaced at reasonable intervals (~one every 5 minutes) throughout the survey. Use attention checks to filter problematic responses (Goodman et al., 2013, p. 20; Jia et al., 2017; Peer et al., 2017; Thomas & Clifford, 2017). |
| | Attention check questions should not require specialized knowledge or information from different survey sections (Downs et al., 2010). |
| Time | Track the time participants take to complete the survey; reject responses or participants that average less than 2 seconds per item. (Wood et al., 2017). |
| Sample Size | Plan to reject a sizeable portion (20% - 50% depending on payment and ex-ante filtering) of responses; adjust the collected sample size accordingly (Aguinis et al., 2021). |
| Participant Payment | Payment affects both the quality of responses and the collection rate. Appropriate pay results in faster, higher-quality responses (Buhrmester et al., 2011; Jia et al., 2017). |
| | Pay a fair, ethical rate. (e.g., fair minimum wage). Platforms often provide guidelines. Worker communities (for example, https://turkopticon.net/) and simple online searches will reveal workers' expectations and inform your conscience (Aguinis et al., 2021; Buhrmester et al., 2011; Lowry et al., 2016). |
| | If a study includes complex or demanding items, pay more, or include bonus payments (Keith et al., 2019). |
| | Formulate precise, objective requirements for participant removal and non-payment. Communicate these requirements to participants at the beginning of the study (Aguinis et al., 2021; Harms & DeSimone, 2015). |
| Participant Pre-Screening | To the extent possible, screen participants to match the intended sample population and context (Buhrmester et al., 2011; Harms & DeSimone, 2015; Jia et al., 2017; Keith et al., 2019; Zhu et al., 2015). |
| | Employ platforms that facilitate ex-ante filtering of participants. (e.g., CloudResearch, Prolific.co) (Lowry et al., 2016). |
| | Filter participants by quality or endorsements; use only those with ≥ 95% ratings(Jia et al., 2017; Steelman et al., 2014). |
| | Filter participants by appropriate location, appropriate language ability(ies), demographic factors, and when appropriate by religion, ideology, or political allegiance (Aguinis et al., 2021; Jia et al., 2017; Lowry et al., 2016; Steelman et al., 2014). |
| Participant Post Screening | Eliminate responses that violate data quality assessments, like attention check questions, the use of bots, and response times. Meticulous rejection of bad responses will increase a sample's legitimacy. (Jia et al., 2017; Steelman et al., 2014; Thomas & Clifford, 2017). |
| | Filter automated responses using mechanisms such as CAPTCHA (Buchheit et al., 2018). |
| | Record and keep the IDs of participants that failed quality checks (Aguinis et al., 2021; Jia et al., 2017). |

There is also agreement in the literature that criticisms of the validity of a study should focus on the composition of the sample, not the means used to collect data (Harms & DeSimone, 2015; Landers & Behrend, 2015; Lowry et al., 2016; Zhu et al., 2015). It is not appropriate to criticize a study *simply* because data was collected using an OCM (Jia et al., 2017). Rather, to ensure data validity, focus on the sample's representativeness. While Mechanical Turk samples are diverse, they do not necessarily represent the global worker population. The researcher must filter and clean the data to ensure its validity. Further, because of the diversity of MTurk samples, we advocate for all the reporting recommendations summarized in Table 17. Documenting data collection procedures is crucial for researchers to ensure the validity and replicability of their data.

**Table 17. OCM Reporting Recommendations (adapted from Steelman et al., 2014)**

1. Ex-Ante Participant Restrictions
   a. Location (i.e., country, state, province, or urban versus rural, etc.)
   b. Language
   c. Technology requirements/restrictions (PC, mobile, etc.)
   d. Survey experience
   e. Quality/approval rating (i.e., ≥ 95%)
   f. Participant demographics:
      i. Age
      ii. Gender
      iii. Race, ethnicity
      iv. Marital status
      v. Income
      vi. Employment status
      vii. Religion
      viii. Ideology
      ix. Political allegiances
2. Payment incentives
   a. Payment
   b. Bonuses
   c. Justification/rationale
3. Task timeline:
   a. Average completion time
   b. Minimum and maximum completion time
4. Description of data quality questions and checks:
   a. Human verification (i.e., CAPTCHA)
   b. Attention checks—Number and form
   c. Description of acceptable and failed/invalid responses
   d. Filtering of previous respondents:
   e. Description of edits or adaptations of scales implemented to improve neutrality and reduce bias
5. Detailed Ex-Post filtering/cleaning
   a. The number of responses pre-filtering
   b. The number of responses/participants excluded for violating data quality assessments
      i. violated time restrictions
      ii. failed attention checks
      iii. participants from previous studies
      iv. Failed verification of (screened to the extent possible):
         1. Demographics
         2. Location
         3. Language
         4. Final sample size
6. Measurement invariance for studies with multiple sample collections
   a. Proof of invariance for combining samples (i.e., checks for unobserved heterogeneity)
   b. OR control for sample groups in subsequent analysis
   c. OR keep as separate samples
7. Generalizability:
   a. The fit between the study sample and the population of interest. (e.g., demographic comparisons)
   b. Acknowledge where sample characteristics differ from the population of interest

Further, we suggest using OCM toolkits like Prolific.co and CloudResearch to filter participants by demographics, geography, quality, and language. OCM samples are diverse, so it is up to researchers to ensure they align with the context of the study. More filtering options can help researchers choose appropriate samples and prevent biases. Most, if not all, of the concerns about the validity and applicability of OCM samples from Table 15 are addressable through the combination of ex-ante filtering and ex-post cleaning. In this study (as in Steelman et al. (2014)), we minimally cleaned our data but believe that further post-hoc cleaning would improve the psychometrics and possibly allow the combination (pooling) of two or more samples.

Microtask workers can exploit the lack of behavioral controls to cheat by lying, using bots, or faking their locations. Detecting cheating on OCM platforms alone is difficult, but OCM toolkits provide extra controls. OCM platforms connect workers through worker ID numbers, allowing for task assignment and compensation. However, these platforms do not verify worker demographics or guarantee satisfactory task completion. Figure 1 outlines a framework for IS tools and their controls.

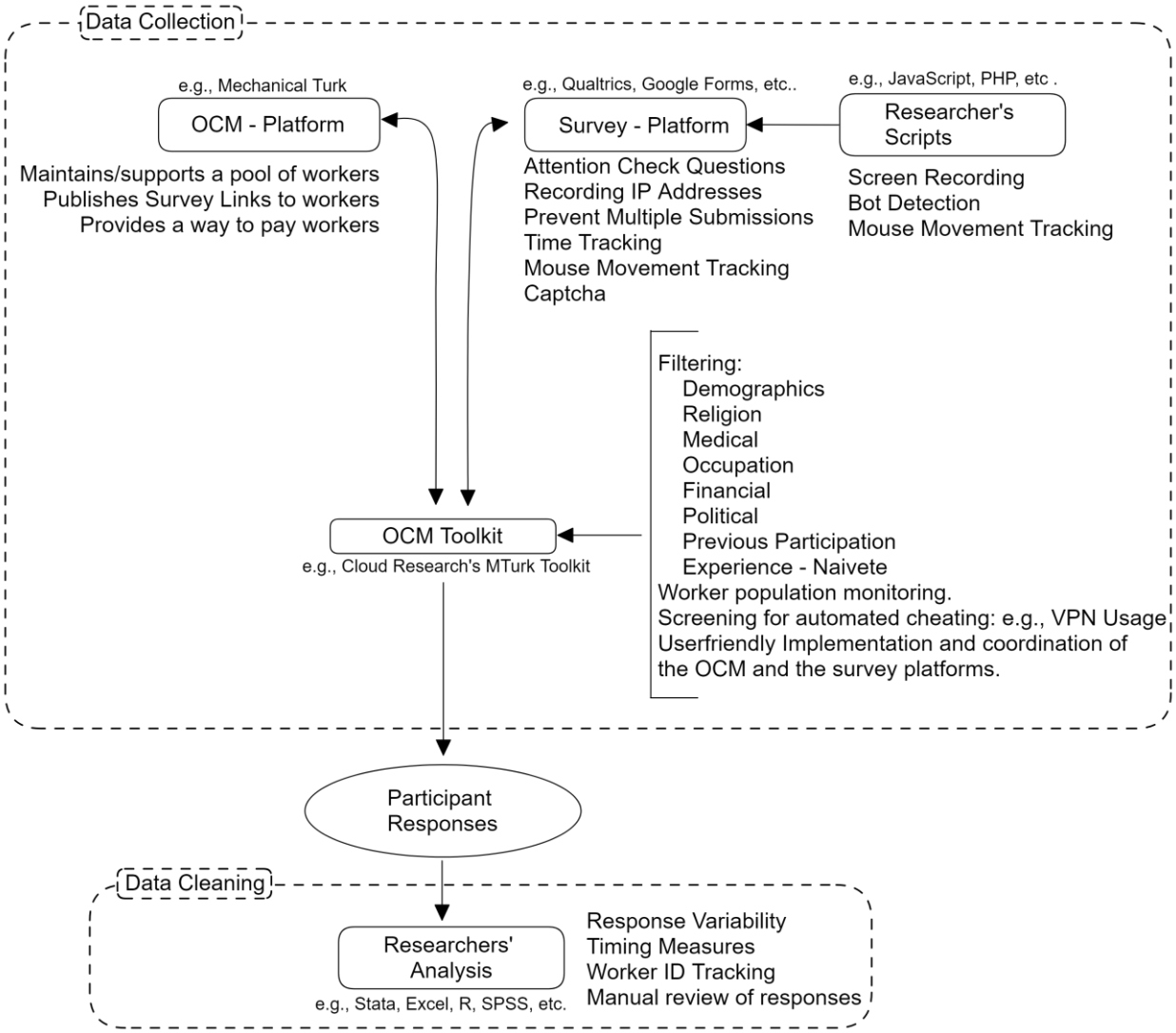


**Figure 1. Tiers of Survey Study Tools**

Online survey platforms have measures to prevent fraudulent responses, such as recording IP addresses, using CAPTCHAs, and preventing multiple submissions. Some platforms offer customizable controls for added security. Above this layer lie the OCM toolkit platforms. OCM toolkit platforms offer user-friendly code-free screening and filtering for unsuitable participants and bad actors in research. Like the CloudResearch MTurk toolkit used for this study[9], toolkit features include participant filtering, worker monitoring, and detecting certain forms of cheating (e.g., VPN location spoofing). The final layer of the hierarchy encompasses the researchers' post-hoc statistical tools and critical thinking skills to test for straight lining, completion time, failed attention checks, and other indications of questionable responses. They may also collect the worker IDs of problematic and cooperative participants.

Researchers should be cognizant of the dependencies embedded in the hierarchy of the tools shown in Figure 1. For example, a researcher may face difficulties in analyzing the time taken by participants to complete a survey task if they forget to include timers on the survey platform. The quality of the data the researcher receives depends on orchestrating the levels in the hierarchy, which depends heavily on the OCM toolkit. Thus, our recommendation for their use.

In 2014, Steelman and colleagues developed guidelines for reporting data obtained from Mechanical Turk and other online crowdsourcing platforms. In addition, we suggest that researchers should test their

[9] Other OCM toolkits, such as Prolific.co or CR Connect, combine the OCM marketplace and toolkit functionalities into one platform.

samples for measurement invariance and report their findings, particularly when gathering a diverse global sample. Alternatively, researchers can clearly show the measures taken to ensure consistency in their results and conclusions drawn from various samples, such as using dummy variables or subset analyses. Like Steelman et al. (2014), we found none of the samples collected was invariant across all measures. Had our focus been consumer use and acceptance of IT, our results and conclusions would have many limitations related to generalizability[10].

Finally, we echo the advice that researchers (Jia et al., 2017; Zhu et al., 2015) should not focus only on the limitations of OCM samples, but also on the unique benefits they can provide when the researchers have designed their study and sample collection appropriately. Valid results require careful selection, analysis, and recording of sample attributes. Additionally, invariant samples can help uncover insights into the boundary conditions of the theory tested, revealing previously unexplored factors.

OCM samples are neither better nor worse than other sampling methods. Researchers should approach OCM samples in a balanced way, recognizing both the opportunities and the diligence costs they present. Reviewers should neither reflexively accept nor reject studies with OCM samples. Reviewers should consider context, research questions, and generalizability when evaluating data suitability. Justifying the sampling pool is crucial for research validity. For example, studies using Mechanical Turk with low compensation that claim to have IT executive respondents should draw scrutiny.

# 5 Limitations and Future Research

Though OCM samples have many advantages, they also have limitations. Concerns about convenience sampling, super-user bias, attention lapses, and contextualization challenges have been raised and addressed in extant literature (e.g., Lowry et al., 2016). Other less addressed issues include ethical concerns for marginalized MTurk workers (Deng et al., 2016). Others caution that anonymity prohibits researchers from knowing whether they are unintentionally employing protected groups, such as minors or prison inmates (Shank, 2016). Participants may also willingly misrepresent themselves to qualify for more exclusive jobs (e.g., those requiring workers to have a specific job title, income, or education level) (Aguinis et al., 2021).

We collected data in 2021 during a COVID-19 lockdown. Students were solicited remotely and completed the survey online at their convenience, like Steelman et al. (2014). However, Steelman et al. (2014) approached students in person. As a student sample moves along a spectrum from entirely in-person to semi-remote (Steelman et al., 2014) to fully remote (this study), participation becomes increasingly digitally mediated. More research is needed to understand how digital mediation affects survey responses, particularly attention lapses and contextualization challenges.

Digital mediation does not explain all the variance in our survey results, differences in the samples matter, too. As with any sample, researchers should consider the demographics of OCM participants carefully before speculating on the generalizability of a study. US college student samples often overrepresent Caucasian Americans ages 18-23 with little or no work experience and an above-average academic inclination. In contrast, OCM workers may be more diverse regarding race, nationality, age, work experience, and education level. Using OCM samples can help researchers overcome some disadvantages of student samples, including the lack of diversity. Further, OCM sample responses were faster and more convenient to collect because we did not have to ask faculty for favors or interrupt classes.

Another limitation of this study is the size of the student sample. During the COVID-19 pandemic, remote work and survey overload reduced easy access to students. This study's limited student sample size illustrates the implications for power. According to the conservative methods for analyzing power in PLS-SEM from Kock and Hadaya (2018), the student model can only detect statistical significance for path coefficients greater than 0.22. The same calculation shows that the OCM samples had sufficient power to detect statistical significance in path coefficients between 0.11 and 0.12[11]. The student sample required an effect size roughly twice that of the OCM samples to achieve the same power level.

[10] For example, the non-U.S., and student samples were not compositionally invariant for the habit construct; the weights (and loadings) creating the habit construct differ significantly between the samples. Because these items combine differently across samples, a meaningful comparison of the habit construct between these populations is not possible with this data.
[11] Assuming (1-β) = .80 and α =.05

While student samples were difficult to secure during the pandemic, Mechanical Turk participation grew as more individuals were out of work and working from home. OCMs helped us collect a larger sample from a wider pool of more willing participants. Gathering equivalent amounts of data from students can require multiple phases. Collecting data from students in waves can be time-consuming. Furthermore, combining these data requires establishing measurement invariance. In contrast, OCMs allow extensive data collection in one wave with timely results.

Finally, our payment of $1, although higher than the 20 cents paid by Steelman et al. (2014), may not be considered adequate in today's crowdsourcing markets. The equivalent hourly rate ranged from $2.93 to $3.40 based on mean or median completion times. Given the changing expectations of workers regarding reasonable pay, some researchers suggest paying the US minimum wage, and platforms like Prolific.co enforce higher payment standards (Aguinis et al., 2021; Prolific Team, 2023). While increasing payment might improve the number of complete OCM responses (Keith et al., 2017), the impact on data quality and speed remains uncertain. Further research on this topic could provide valuable insights for researchers regarding the effect of payment on completion rates and response quality.

# 6 Conclusion

This research confirms the findings of Steelman et al. (2014) supporting the use of Mechanical Turk (MTurk) samples as alternatives to student samples, particularly for individual-level research on general consumer populations. Our partial replication compared MTurk workers with students, focusing on a virtual-reality headset as the IT artifact. We observed that MTurk samples responded more similarly to each other than students, and the data from MTurk workers aligned better with well-established theoretical relationships in UTAUT 2. Based on our findings, we provide recommendations for researchers collecting and reporting data from MTurk and for reviewers evaluating studies involving MTurk samples.

## References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535.

Belkin, D. (2021, September 6). *A generation of american men give up on college: 'I just feel lost.'* WSJ. Retrieved from https://www.wsj.com/articles/college-university-fall-higher-education-men-women-enrollment-admissions-back-to-school-11630948233

Buchheit, S., Doxey, M. M., Pollard, T., & Stinson, S. R. (2018). A technical guide to using Amazon's Mechanical Turk in behavioral accounting research. *Behavioral Research in Accounting*, *30*(1), 111–122.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Chin, W. W. (2002). Partial least squares for IS researchers: An overview and presentation of recent advances using the PLS approach. *ICIS 2000 Proceedings*, *88*. Retrieved from http://aisel.aisnet.org/icis2000/88

Daly, T. M., & Nataraajan, R. (2015). Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk. *Journal of Business Research*, *68*(12), 2603–2609.

Deng, X. (Nancy), Joshi, K. D., & Galliers, R. D. (2016). The duality of empowerment and marginalization in microtask crowdsourcing: Giving voice to the less powerful through value sensitive design. *MIS Quarterly*, *40*(2), 279–302.

Dennis, A., & Valacich, J. (2014). A replication manifesto. *AIS Transactions on Replication Research*, *1*, 1–4.

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 135–143).

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system?: Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 4.

Duffin, E. (2021, June 11). *Americans with a college degree 1940-2018, by gender*. Statista. Retrieved from https://www.statista.com/statistics/184272/educational-attainment-of-college-diploma-or-higher-by-gender/

Facebook. (2021). *Oculus Quest 2: Our most advanced new all-in-one VR headset*. Oculus. Retrieved from https://www.oculus.com/quest-2/

Fang, J., Prybutok, V., & Wen, C. (2016). Shirking behavior and socially desirable responding in online surveys: A cross-cultural study comparing Chinese and American samples. *Computers in Human Behavior*, *54*, 310–317.

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, *75*, 47–52.

Ford, J. B. (2017). Amazon's Mechanical Turk: A comment. *Journal of Advertising*, *46*(1), 156–158.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, *19*(4), 440–452.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, *18*(3), 382–388.

Georgetown University. (2021, September 10). *Women increasingly outnumber men at US colleges—But why?* THE FEED. Retrieved from https://feed.georgetown.edu/access-affordability/women-increasingly-outnumber-men-at-u-s-colleges-but-why/

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling [PLS-SEM]* (3rd ed.). SAGE Publications.

Hair, J. F., Jr., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1), 2–24.

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology*, *8*(2), 183–190.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 2158244015584617.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International Marketing Review*, *33*(3), 405–431.

Jia, R., Steelman, Z. R., & Jia, H. H. (2022). What makes one intrinsically interested in IT? An exploratory study on influences of autistic tendency and gender in the US and India. *MIS Quarterly*, *46*(3), 1603–1634.

Jia, R., Steelman, Z., & Reich, B. H. (2017). Using Mechanical Turk data in IS research: Risks, rewards, and recommendations. *Communications of the Association for Information Systems*, *41*, 301–318.

Jiang, L., Wagner, C., & Chen, X. (2021). Taking time into account: Understanding microworkers' continued participation in microtasks. *Journal of the Association for Information Systems*, *22*(4), 893-930.

Keith, M. G., Harms, P., & Tay, L. (2019). Mechanical turk and the gig economy: Exploring differences between gig workers. *Journal of Managerial Psychology*, *34*(4), 286–306.

Keith, M. G., Tay, L., & Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, *8*, 1359.

Kock, N. (2015). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of E-Collaboration*, *11*(4), 1–10.

Kock, N. (2020). Harman ' s single factor test in PLS-SEM : Checking for common method bias. *Data Analysis Perspectives Journal*, *2*(April), 1–6.

Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM : The inverse square root and gamma-exponential methods. *Information Systems Journal*, *28*(1), 227–261.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142–164.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442.

Lowry, P. B., D'Arcy, J., Hammer, B., & Moody, G. D. (2016). "Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels. *The Journal of Strategic Information Systems*, *25*(3), 232–240.

Lowry, P. B., Moody, G. D., & Chatterjee, S. (2017). Using IT Design to Prevent Cyberbullying. *Journal of Management Information Systems*, *34*(3), 863–901.

Lowry, P. B., & Wilson, D. (2016). Creating agile organizations through IT: The influence of internal IT service perceptions on IT service quality and IT agility. *The Journal of Strategic Information Systems*, *25*(3), 211–226.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Prolific Team. (2023, February 3). *Prolific's payment principles*. Prolific. Retrieved from https://researcher-help.prolific.co/hc/en-gb/articles/4407695146002-Prolific-s-payment-principles

Reeves, R. V., & Smith, E. (2021, October 8). The male college crisis is not just in enrollment, but completion. *Brookings*. Retrieved from https://www.brookings.edu/blog/up-front/2021/10/08/the-male-college-crisis-is-not-just-in-enrollment-but-completion/

Ringle, C. M., Wende, S., & Becker, J.-M. (2015). *SmartPLS 3*. SmartPLS.

Ringle, C. M., Wende, S., & Becker, J.-M. (2022). *SmartPLS 4*. SmartPLS.

Shank, D. B. (2016). Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk. *The American Sociologist*, *47*(1), 47–55.

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age: Innovative alternatives to student samples. *MIS Quarterly*, *38*(2), 355–378.

Tamilmani, K., Rana, N. P., & Dwivedi, Y. K. (2021). Consumer acceptance and use of information technology: A meta-analytic evaluation of UTAUT2. *Information Systems Frontiers*, *23*(4), 987–1005.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425.

Venkatesh, Thong, J. Y. L., & Xu. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, *17*(5), 328–376.

Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2019). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology*, *34*(4), 425–452.

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464.

Zhu, X. (Susan), Barnes-Farrell, J. L., & Dalal, D. K. (2015). Stop apologizing for your samples, start embracing them. *Industrial and Organizational Psychology*, *8*(2), 228–232.

## About the Authors

**Stephen A. De Lurgio II** is a Ph.D. student in Information Systems in the Sam M. Walton College of Business at the University of Arkansas. He holds a BS in Aerospace Engineering from the University of Kansas and an MBA from St. Louis University. His research interests are emerging technologies, their application, and impacts.

**Amber Young** is the Director of the Information Systems Ph.D. Program and Associate Professor of Information Systems in the Sam M. Walton College of Business at the University of Arkansas. Her current research focuses on how information systems design can promote social good and positive organizational outcomes. Amber serves as Associate Editor for *MIS Quarterly* and is on the editorial board of *Information & Organization*. Her research appears in *MIS Quarterly* (x2), the *Journal of Management Information Systems*, *MIT Sloan Management Review*, *Journal of the AIS*, *Information Systems Journal*, *Information & Organization*, *International Journal of Information Management,* and *Communications of the AIS*.

**Zachary R. Steelman** is an Associate Professor of Information Systems in the Walton College of Business at the University of Arkansas. He has authored refereed publications in prominent IS publications such as *Information Systems Research*, *MIS Quarterly*, *MIS Quarterly Executive*, *Information Systems Journal*, *Communications of the Association of Information Systems*, *Americas Conference on Information Systems*, and the *Hawaii International Conference on System Sciences*. He was recently awarded the 2020 AIS Early Career Award which “recognizes individuals in the early stages of their careers who have already made outstanding research, teaching, and/or service contributions to the field of information systems”. He has had recent publications in the area of neurodiversity in *MIS Quarterly*, conferences, and workshops as well as Autism-specific journals and books. He has recently worked with multiple Fortune 500 organizations examining the impact of IT and neurodiversity policy changes on individuals and the organization.