Published in final edited form as:

Comput Human Behav. 2019 May ; 94: 1–8. doi:10.1016/j.chb.2018.12.042.

MTurk Participants Have Substantially Lower Evaluative Subjective Well-Being Than Other Survey Participants

Arthur A. Stone^{a,b}, Marta Walentynowicz^a, Stefan Schneider^a, Doerte U. Junghaenel^a, and Cheng K. Wen^a

^aDornsife Center for Self-Report Science, University of Southern California, 635 Downey Way Los Angeles, CA 90089, United States.

^bDepartment of Psychology, University of Southern California, 3620 South McClintock Ave., Los Angeles, CA 90089, United States.

Abstract

Amazon's MTurk platform has become a popular site for obtaining relatively inexpensive and convenient adult samples for use in behavioral research. Concerns have been raised about selection issues, because MTurk workers chose to participate in the platform and select the tasks they perform (of many offered to them). Prior studies have documented demographic and psychological differences with national samples. In this paper we studied evaluative subjective well-being (the Cantril Ladder) in an MTurk sample, a national Internet panel sample, and a national telephone survey conducted by Gallup-Sharecare. A surprising finding was that MTurk participants' Ladder scores were substantially lower than the other two samples. Analyses controlling for six demographic differences among the samples only slightly reduced the mean differences. However, patterns of demographic—well-being associations were similar within the samples. To corroborate these results, we conducted a secondary analysis on another three samples, one MTurk sample and two Internet panel samples. The same group differences in Ladder scores were observed. These findings add to the growing literature documenting the characteristics of MTurk samples and we discuss the implications for future research with such samples.

Keywords

MTurk; Cantril ladder; Subjective Well-being

Corresponding author: Marta Walentynowicz, PhD, Dornsife Center for Self-Report Science, & Center for Economic and Social Research, University of Southern California, Suite 405 635 Downey Way, Los Angeles, CA 90089, Tel: (213) 821-8854, marta.walentynowicz@gmail.com.

Declaration of Interest:

Arthur A. Stone is a Senior Scientist with the Gallup Organization and a consultant with Adelphi Values, inc.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Much has been written about the potential applicability of using Amazon Mechanical Turk (MTurk) participants for studying social and behavioral phenomena (Chandler & Shapiro, 2016; Keith, Tay, & Harms, 2017; Shank, 2016; Shapiro, Chandler, & Mueller, 2013; Smith, Sabat, Martinez, Weaver, & Xu, 2015), which is understandable considering the relatively low cost and expeditiousness of data collection with this system. Researchers now have the capability of running low-cost survey studies and having the data available in a matter of days, making such studies particularly attractive and highly popular (Anderson et al., 2018). On the downside, there is the likelihood that samples drawn from the MTurk population, which has more than half a million registered users according to Amazon, are not demographically representative of the US population as a whole. A number of studies showed that MTurk samples are different from samples that are representative of the US population on several demographic characteristics (e.g., younger and better educated, but with lower income and a higher proportion of workers with European- and Asian-American ethnic background) (Berinsky, Huber, & Lenz, 2012; Keith et al., 2017; McCredie & Morey, 2018; Paolacci & Chandler, 2014; S. M. Smith, Roster, Golden, & Albaum, 2016; Walters, Christakis, & Wright, 2018) and psychological characteristics (e.g. more cognitive symptoms and more likely to be depressed, anxious, and socially isolated) (Arditte, Çek, Shaw, & Timpano, 2016; McCredie & Morey, 2018; Walters et al., 2018). Other studies have shown that MTurk participants have another characteristic: they are more attentive to the instructions for surveys than convenience samples, such as undergraduate students, perhaps because the typical MTurk tasks “teach” the individuals to carefully consider instructions (Hauser & Schwarz, 2016; Ramsey, Thompson, McKenzie, & Rosenbaum, 2016). There is certainly much more to be learned about MTurk participants going beyond demographic, psychological, health, and attention differences that could either increase or decrease confidence in the external validity of results based on such samples and in the appropriateness of using MTurk samples for particular research questions.

In this spirit, we report a startling finding from a recent study where self-reports of evaluative subjective well-being were collected from both MTurk and other panel samples (reference omitted for the purpose of double-blind review, hereafter referred to as Stone et al., 2018). The measure of subjective well-being was the Cantril’s Self-Anchoring Scale (henceforth called the “Ladder,” Cantril, 1965), a single-item question that has been frequently used in studies tracking the subjective well-being of nations (Steptoe, Deaton, & Stone, 2015; Stone, Schwartz, Broderick, & Deaton, 2010). Subjective well-being is increasingly recognized as an important measurement for assessing quality of life (Steptoe et al., 2015) and health evaluation (Dolan & White, 2007). Comparing subjective well-being from both MTurk and other representative panel samples can, therefore, provide evidence for the utility of MTurk for studies interested in evaluating subjective well-being and for those that could be affected by subjective well-being. To foreshadow the results, the surprising finding was that Ladder scores in the MTurk sample were *much* lower than in nationally representative samples (e.g., Deaton, 2011; Deaton & Stone, 2016).

Our goal in this paper is to document differences in subjective well-being measured by the Ladder among the MTurk and other national samples and to ascertain if they were caused by

the demographic differences between those samples. If we find that controlling for demographic variables eliminates differences in the Ladder among the samples, then we can feel more secure that estimates obtained from studies with MTurk samples are representative of the general population if sample weighting procedures are employed. However, if substantial Ladder differences remain after demographic controls are included in the models, then this raises potential concerns about the use of MTurk samples for examining topics associated with subjective well-being and possibly other topics. To address this research question, we initially compared the Ladder scores in three samples. Data for two samples were retrieved from a previous study (Stone et al., 2018) and included participants from MTurk and from a nationally representative Internet panel. The third sample was retrieved from the Gallup-Sharecare daily survey that collected data from a very large US sample over the years 2014–2016. Because all studies included both the Ladder and demographic data, we were able to test the hypothesis that observed differences in self-reported well-being levels were due to demographic differences between the samples. Finally, we performed a literature search for other studies measuring subjective well-being with the Ladder in the MTurk population. We were able to locate two studies (Busseri & Samani, 2018; Whillans, Dunn, Smeets, Bekkers, & Norton, 2017), one of which had administered the Ladder in MTurk and two other samples (Whillans et al., 2017), allowing for a replication of the difference in Ladder scores.

2. Methods

2.1. Participants and Procedure

2.1.1. MTurk sample—The data from this sample were collected as part of a study designed to examine item context effects (Schwarz, 1999) on the measure of subjective well-being and potential manipulations to reduce the context effects. For the complete description of the study methods, see Stone et al., 2018. Participants ($n=4,500$) were recruited through the MTurk website. The study was open to respondents above 18 years of age who were located in the United States, completed at least 100 MTurk tasks, and had high task completion rates (called approval rates) for their previous MTurk tasks ($> 95\%$). Participants completed a 4-minute survey which included a measure of subjective well-being, a political question asking about the direction that the country is going, and a set of demographic questions. The order of question presentation (9 conditions) was counterbalanced across participants. Participants who failed quality check questions (i.e., a very easy question that everyone should answer correctly if they were paying attention to the task) were excluded from the analyses ($n=233$, 5.2%). Participants were compensated \$0.45 for their participation (a low subject payment typical in MTurk studies). The study was approved by the local Institutional Review Board. For the purpose of the present analyses, only participants who answered the Ladder as the first question in the survey were retained ($n=933$; data from other participants were not used here to avoid confounding influences of experimental conditions).

2.1.2. UAS Internet Panel sample—Data from this sample were collected in a study aimed to replicate the findings from the MTurk sample regarding the item context effects on the Ladder. The complete description of the study methods is described in Stone et al., 2018.

Respondents were recruited from the Understanding America Study (UAS), a probability-based Internet panel of approximately 6,100 individuals from across the United States. The panel is hosted by the Center for Economic and Social Research at the University of Southern California (<https://uasdata.usc.edu>). For this study, invitations were sent to 6,094 panelists, out of which 4,579 completed the survey (response rate of 75.1%). The study was embedded in a larger survey consisting of 5 question blocks. In our block, participants were randomized into four conditions based on various combinations of the Ladder and political question. Participants were paid \$3 for completing the study. The UAS surveys, recruiting, and consent procedures were reviewed by the local Institutional Review Board. As for the MTurk sample, only participants who received the Ladder as the first question in the survey were included in the present analyses (n=221).

2.1.3. Gallup-Sharecare sample—Gallup-Sharecare (<https://wellbeingindex.sharecare.com>) is a nationally representative telephone survey that interviews at least 500 U.S. adults aged 18 and older daily throughout the year. It uses a dual-frame random-digit dial methodology including both landline and cellphone numbers from the 50 states in the United States and the District of Columbia. Each sample of national adults includes a minimum quota of 70% cellphone respondents and 30% landline respondents. The survey includes items assessing many aspects of well-being. For the present analyses, we retrieved data collected from January 2014 to December 2016 that included complete demographic information and the Ladder (n=137,185).

2.1.4. Additional samples—To examine the replicability of MTurk versus other samples result, we reanalyzed data from Whillans et al. (2017). This study collected Ladder scores from an MTurk sample (n=366), a sample from the GFK Knowledge Networks Panel (n=1,260), and a Qualtrics sample (n=1,802). These data were used to replicate mean differences in Ladder scores across MTurk and other samples.

2.2. Measures

2.2.1. Ladder—Cantril's Self-Anchoring Scale (Cantril, 1965) was used to measure evaluative subjective well-being. It is a single-item question which requests individuals to position themselves on a 11-step ladder with 0 representing the worst possible life and 10 representing the best possible life. The presentation of Ladder question in the current study is shown in Figure 1. In the study of Whillans et al. (2017), the response options were also presented vertically, but with 0 at the top and 10 at the bottom. In the Gallup telephone survey, the options were verbally presented to participants.

2.3 Analytic Plan

2.3.1. Harmonization of demographic variables.—Demographic variables are the primary predictors in the present analyses. Although questions measuring demographics were to some extent similar across the samples, the response categories differed. A common set of response options needed to be created for analyses of the combined samples. Table 1 presents the list of the response categories for the harmonized demographic variables in each sample. A harmonized education variable was coded as follows: Less than high school, High school graduate, Some college, College graduate, and Post graduate. A harmonized race

variable comprised the categories White, Black, and Other, but we note that for Gallup-Sharecare coding, the definition of White and Black include only non-Hispanic, whereas that is not the case for the MTurk/UAS Internet Panel studies. In case of income, categories used by the Gallup-Sharecare sample were not easily collapsible with the MTurk/UAS Internet Panel studies (see Table 1). A harmonized income variable had the following categories: < \$20,000, \$20,000 - \$34,999, \$35,000 - \$49,999, \$50,000-\$74,999, and > \$75,000; the exact cut-points for the income categories differed somewhat in the Gallup-Sharecare sample and the category labels are based on the MTurk/UAS Internet Panel studies (although the discord in this harmonization could have been minimized by reducing the number of categories, we decided against this alternative in order to maintain more fine-grained information). A harmonized marital status variable followed the categories of the MTurk sample which included: Never married, Married, Living with partner, Separated, Divorced, and Widowed. In addition, continuous age was categorized into 6 bins, 20–29, 30–39, 40–49, 50–59, 60–69, and 70–80. The gender variable included Male and Female categories in all samples, thus did not have to be harmonized.

2.3.2. Analysis—Univariate ANOVA was used to determine if there were differences in Ladder scores between the three samples. Subsequent ANCOVA models added the set demographic covariates to the model to determine if the differences in Ladder scores between the samples persisted. Additional analyses added the interaction terms between sample and demographic variables to examine whether sample differences were consistent across the levels of each demographic variable. Graphical examination of the Ladder scores by sample and selected demographic variables were used to visualize resulting patterns. All analyses were conducted in STATA.

3. Results

3.1. MTurk, UAS Internet Panel, and Gallup-Sharecare Ladder comparison

We had three samples – the MTurk and UAS Internet Panel samples and the Gallup-Sharecare sample – for testing Ladder differences and for determining if demographic differences explained observed differences. Table 2 presents the mean Ladder score (and standard deviation) and demographic characteristics in each of the three samples. The Ladder score differed significantly by sample, $F(2, 138336)=447.6$, $p<.0001$. Post hoc pairwise comparisons showed that the MTurk sample differed from the UAS Internet Panel and Gallup-Sharecare samples, whereas the latter two samples did not significantly differ from each other. The magnitude of the differences in standardized effect sizes was Cohen's $d = .83$ for the MTurk-UAS comparison and $d = .98$ for the MTurk-Gallup comparison (considered large effects following Cohen's conventions). Regarding the demographic variables, all of the demographic variables differed between the three samples: Age (categorical), $\chi^2(10)=824.8$, $p<.001$, Gender, $\chi^2(2)=46.9$, $p<.0001$, Education, $\chi^2(8)=251.7$, $p<.0001$, Marital status, $\chi^2(10)=343.1$, $p<.001$, Race, $\chi^2(4)=14.1$, $p<.007$, and Income, $\chi^2(8)=186.9$, $p<.0001$.

As these expected demographic differences could be confounding the differences in the Ladder scores between the samples, the next step was to statistically adjust the Ladder

means for these demographic variables. A consideration in these analyses is that some of the demographics had considerable missing data (e.g., Income). We re-ran the first ANOVA model, which did not include the control variables, using only individuals who had complete data on all of the demographic variables ($n=138,339$) to ensure that a comparison of models with and without covariates was not biased by examining different samples (Table 3, Model 1). The second ANOVA included the demographic control variables (Table 3, Model 2). For Model 1, the mean Ladder scores in the UAS Internet Panel and the Gallup-Sharecare samples were 1.66 and 1.86 points higher, respectively, compared to the MTurk sample. The same mean differences in Model 2 were slightly reduced, 1.50 and 1.65, though they continued to be highly significant, $F(2, 138,314)=394.7$, $p<.0001$. We tested the reduction in the mean differences using STATA's SUEST procedure (Seemingly Unrelated Estimation), which allows a comparison of coefficients from two different models with overlapping data. The reduction was significant for the comparison between the MTurk and the UAS Internet Panel samples, $\chi^2(1)=10.65$, $p<.01$, as well as for the Gallup-Sharecare sample, $\chi^2(1)=100.1$, $p<.001$. With regard to the effects of the demographic variables on the Ladder, all variables were significantly associated with it: Age, $F(5, 138,314)=387.3$, $p<.0001$, Gender, $F(1, 138,314)=699.9$, $p<.0001$, Education, $F(4, 138,314)=164.8$, $p<.0001$, Race, $F(2, 138,314)=195.1$, $p<.0001$, Marital Status, $F(5, 138,314)=206.9$, $p<.0001$, and Income, $F(4, 138,314)=1580.4$, $p<.0001$.

Although the finding of a lower average Ladder score in the MTurk sample compared to the other two samples was maintained after adjusting for demographic controls, we wished to determine if the differences between samples were consistent over levels of the demographic variables. For example, MTurk Ladder averages could be more pronounced at one level (e.g., males) than another level (e.g., females) of a variable. To examine this, sample by demographic variable interaction terms were added to the previous ANCOVA model (Model 2), so that each interaction term was controlled for all of the other main effect and interaction terms. Of the 6 interaction terms, 2 were significant at the .01 alpha level: Sample $\times$ Race, $F(4, 138,272)=3.77$, $p<.01$ and Sample $\times$ Income, $F(8, 138,272)=4.68$, $p<.001$. Visual inspection revealed very little differences in the Sample $\times$ Race pattern of results, but the Sample $\times$ Income showed a wider gap between MTurk and the remaining two groups at lower income levels (about 2 points on the Ladder) versus at higher income levels (about 1 point; see Figure 2).

To better understand these interactions, the same analyses were performed by pairs of groups formed by the grouping variable (MTurk/Internet Panel/Gallup-Sharecare), that is, MTurk v. Internet Panel, MTurk v. Gallup-Sharecare, and Internet Panel v. Gallup-Sharecare. For Income as the predictor variable, the interaction terms for the three analyses were as follows: MTurk v. Internet Panel, $F(4, 1109)=1.22$, $p>.05$, MTurk v. Gallup-Sharecare, $F(4, 138074)=8.29$, $p<.0001$, and, Internet Panel v. Gallup-Sharecare, $F(4, 137361)=1.09$, $p>.05$. Thus, the overall interaction effect for Income is apparently coming from the difference between Gallup-Sharecare and MTurk groups, and the slope for the MTurk group was steeper than for the Gallup-Sharecare group. The same procedure was used to explore the Race interaction: MTurk v. Internet Panel, $F(4, 1109)=0.11$, $p>.05$, MTurk v. Gallup-Sharecare, $F(4, 138074)=6.49$, $p<.002$, and, Internet Panel v. Gallup-Sharecare, $F(4, 137361)=1.05$, $p>.05$. The interaction appears to be caused by the relatively larger difference

in group means (MTurk v. Gallup-Sharecare) at the Other race level compared with the differences at the White and Black levels. Of course, these findings are affected by the large N for the Gallup-Sharecare group.

A final analysis to determine Ladder differences among these three samples controlled for possible interactive effects *among* selected demographic variables¹. In part, the work was inspired by Huff and Tingley's (2015) paper examining joint distributions among age, gender, and race in an MTurk sample. We extended this analytic strategy to include income. In order to examine the interactions among the four variables predicting Ladder scores and to keep the number of distinct cells formed by the interaction term reasonable, we treated the age and income variables in these analyses as continuous (whereas in prior analyses we used them as categorical variables). None of the many resulting interaction terms was significant at the .01 level. Most telling, though, was that the Ladder means adjusted for the main effects of the demographic variables were almost the same as those resulting from regression controlling for interactions among the demographic variables and group. The means for the three groups were (main effects adjusted following by interaction adjusted): MTurk, 5.46 v. 5.48; Internet Panel, 6.96 v. 7.05; and, Gallup-Sharecare, 7.11 v. 7.11. These results suggest that interactions among the demographic variables are not responsible for the very low Ladder scores in the MTurk group.

3.2. Reanalysis of a previously published study

The previous analyses clearly show that our MTurk sample had lower evaluative well-being than the comparison samples. There are, though, possible weaknesses to these comparisons that we sought to remedy in a supplemental analysis of a prior study. The first concern is that Gallup-Sharecare sample was from 2014–2016 whereas the MTurk sample was collected in 2018. Perhaps there have been shifts in well-being over time that could explain the Ladder differences. However, this explanation is largely diminished by the UAS Internet Panel sample's findings since these data were also collected in 2018. Second, we rely only on one MTurk sample and it is possible that the study description presented to the MTurk participants resulted in a selection bias. The current study was advertised as one concerning subjective well-being, and it is conceivable that those who were unhappy (low subjective well-being) decided that the task might be of interest, creating a selection bias. It is possible that other MTurk tasks offered with different study descriptions (unrelated to subjective well-being) would not have this selection issue and therefore could show more typical levels of well-being.

To address these concerns, we searched the literature through PubMed, Web of Science, and Google Scholar databases for studies with the following criteria: 1) assessing subjective wellbeing with the Cantril Ladder and 2) among participants from an MTurk population. Two studies met those criteria (Busseri & Samani, 2018; Whillans et al., 2017). The study of Busseri and Samani (2018) investigated lay theories for life satisfaction and reported the following Ladder scores in two MTurk samples: Study 1, N=320, 5.98 (SD=2.27); Study 2,

¹We thank an anonymous reviewer for this suggestion.

N=321, 6.09 (SD=2.15). Those scores are higher than the scores of our MTurk sample, but lower than the scores observed in the UAS and Gallup samples.

A recent study by Whillans and colleagues (2017) not only met the criteria but also included other nationally representative and online samples. This study examined the question of whether or not money and time usage were associated with evaluative well-being. Several samples were examined and three are pertinent to our question: an MTurk sample of 366, a US representative sample obtained from the GfK Knowledge Networks Panel (Internet) of 1,260, and a Qualtrics sample of 1,802. However, due to the fact that the study's focus was not directed at sample differences in subjective well-being, average Ladder scores by sample were not presented in the published article. Fortunately, raw data from this study were available online (Whillans, 2017), allowing us to compute MTurk and comparison sample Ladder means. The average Ladder scores for the three samples were: MTurk, 5.32 (SD=2.05); GfK, 6.76 (SD=1.74); and, Qualtrics, 6.80 (SD=1.99). The means were significantly different, $F(2, 3,466)=97.7, p<.0001$ and post hoc testing revealed that the MTurk mean was significantly lower than the other two samples, and those two samples did not differ from each other. Effect sizes were $d = .79$ for the MTurk-GfK comparison and $d = .74$ for the MTurk-Qualtrics comparison. In addition, the mean of the MTurk sample in the Whillans study was very similar to the mean of the MTurk sample in our primary study described above (a summary of the Ladder means for both studies is shown in Figure 3).

4. Discussion

The motivation for this paper came from an unexpected observation in a recent study showing that a sample of MTurk workers scored surprising low on a measure of evaluative subjective well-being compared to a sample from a more representative Internet panel. Based on our prior experience working with the Cantril Ladder, we thought that the observed group difference of 1.6 rungs/points on the Ladder was exceptionally large. To place this value into context, a difference of 0.93 rungs was observed among those who report being disabled versus those who do not in the Gallup-Sharecare survey (in a sample of over 2,500,000 participants), with those participants reporting disability scoring lower. Our initial thought about this difference was that the typical demographic characteristics of MTurk participants relative to other nationally-drawn samples would explain most of the observed effect. But this was not the case: when Ladder scores were controlled for six demographic characteristics, there was only a slight diminution of the difference and MTurk participants remained remarkably low on the Ladder. Furthermore, controlling the group means for interactions among four of the demographic variables did not alter the finding.

To extend our understanding of the MTurk participants' well-being, we also compared them to a very large sample of individuals surveyed by Gallup-Sharecare a few years ago, which was a telephone interview study (a different mode of Ladder administration). The Ladder score from this sample was very similar to that from our Internet panel and statistically controlling for demographics again did little to reduce group differences. In addition, we were able to reanalyze data from a recent study, which also included a MTurk sample and two Internet panel samples, but that had entirely different aims than ours. The results from this reanalysis replicated MTurk participants' lower position on the Ladder.

While it seems clear that the self-reported subjective well-being differences between MTurk samples and more representative samples are substantial, the relationships between demographic characteristics and scores on the Ladder *within* each of our samples were generally comparable. Of the six demographics we examined, only one, Income, demonstrated a differential association with the Ladder in the MTurk sample versus our UAS Internet Panel and the Gallup-Sharecare samples. In particular, there was a larger gap in Ladder scores at lower versus higher levels of income. Nevertheless, at every level of income MTurk sample had at least a 1-point lower score on the Ladder. And another sample of MTurk participants who completed the Ladder also had scores that were considerably lower than those of the representative Internet panels.

We have several points of discussion about these findings. The first has to do with *why* MTurk participants report such low satisfaction with their lives, though we can only speculate at this point as we do not have data to directly support the possibilities. A very plausible explanation is that our demographic variable list, which was admittedly limited, failed to include other pertinent demographic, psychological, and social variables that would explain the differences, as subjective well-being is also affected by many factors including family relationship, social roles, and health status (Steptoe et al., 2015). It is not clear which other relevant concepts might be proposed for future study, but one can imagine, for instance, that the tasks MTurk participants engage in could be more attractive to individuals who seek high levels of stimulation or who like to always keep busy. Another possibility is that MTurk participants are in more financial debt than others, an issue that might not be reflected in their income levels. It may be that these qualities are also associated with lower well-being and could explain the observed differences. Another line of reasoning centers on the context in which the Ladder question is being completed. Perhaps MTurk participants view responding to questionnaires advertised on the MTurk platform as a work-related event. If the MTurk work is typical of other types of working, then it is probably hedonically negative, as shown by momentary studies (e.g., Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). But to implicate this explanation, one needs to assume that this particular context could, in fact, influence Ladder scores. As the Ladder is supposed to capture a stable judgment rather than a fluctuating, momentary experience, the contextual explanation does not seem very plausible. Conversely, we also know that item context effects can influence Ladder scores as shown in Deaton's recent work (Deaton, 2011). Thus, we view this as a plausible, if not likely, explanation. Finally, it could also be speculated that participants respond to surveys administered via telephone differently than to online surveys (Krosnick & Alwin, 1987). However, both the MTurk and UAS studies were administered online, making this explanation less plausible. Taken together, none of these explanations is entirely satisfying and we hope future studies take this question further.

Yet another method issue that may have influenced our findings concerns the measurement of income and our methods for harmonizing this variable over the three samples. There are issues in the measurement of income having to do with how individuals aggregate income over various income sources and with peoples' willingness to accurately disclose income (Alwin, Zeiser, & Gensimore, 2014; Davern, Rodin, Beebe, & Call, 2005). It is also plausible that the different response options for income reporting in the three surveys used here may have affected the Ladder by sample interactions. We attempted to harmonize the

three sets of response options, but must acknowledge that this was not entirely satisfactory, because the income response options did not always align in a precise way.

The second point of discussion is whether or not our finding of lower well-being of MTurk participants matters for *interpreting findings* when MTurk samples are used in empirical research. On the one hand, the association between demographic variables and the Ladder did not differ much between MTurk and more representative samples (with the exception of income). One could assume that this finding would generalize to other associations between the Ladder and other variables and conclude that researchers interested in subjective well-being, in particular in the associations between different demographic factors and well-being, could rely on MTurk samples without much concern. On the other hand, the observation that the MTurk respondents have considerably lower levels of evaluative subjective well-being could impact other types of research, such as studies focused on recall of negatively valenced events or cognitive processing tasks, which can be influenced by negative affect (Egidi & Caramazza, 2014; Thorley, Dewhurst, Abel, & Knott, 2016).

Finally, we also reported that the income by sample interaction showed a greater spread of the Ladder between those with lower incomes than those with higher incomes. This, too, is curious and may be problematic for MTurk studies examining income or wealth in relation to well-being (where one might draw different conclusions with an MTurk sample). However, we acknowledge that even though we harmonized the variables, the wording of the questions (e.g., income question) differed slightly between the surveys, which could affect the comparability of responses across samples.

In conclusion, this paper contributes to the growing literature on MTurk participants and on evaluative well-being. Our findings clearly indicate that MTurk participants tend to be less satisfied with their lives than more nationally representative samples, but the reasons for this discrepancy and its implications are far from obvious. We hope that this and prior work on MTurk participants inspires new research on this very interesting group of people, one that is increasingly being recruited for social, behavioral, and clinical studies.

Acknowledgments:

This work was supported by a grant from the National Institute on Aging [grant number R01 AG042407; AAS, PI]. The authors thank the Gallup Organization for access to their survey results and the Understanding America Study for conducting the survey.

References

- Alwin DF, Zeiser K, & Gensimore D (2014). Reliability of self-reports of financial data in surveys: Results From the Health and Retirement Study. *Sociological Methods & Research*, 43, 98–136. 10.1177/0049124113507908
- Anderson CA, Allen JJ, Plante C, Quigley-McBride A, Lovett A, & Rokkum JN (2018). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, 1–9. 10.1177/0146167218798821
- Arditte KA, Çek D, Shaw AM, & Timpano KR (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, 28, 684–691. 10.1037/pas0000217 [PubMed: 26302105]

- Berinsky AJ, Huber GA, & Lenz GS (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20, 351–368. 10.1093/pan/mpr057
- Busseri MA, & Samani MN (2018). Lay theories for life satisfaction and the belief that life gets better and better. *Journal of Happiness Studies*. 10.1007/s10902-018-0016-x
- Cantril H (1965). *Pattern of human concerns*. New Brunswick, NJ: Rutgers University Press.
- Chandler J, & Shapiro D (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, 12, 53–81. 10.1146/annurev-clinpsy-021815-093623
- Davern M, Rodin H, Beebe TJ, & Call KT (2005). The effect of income question design in health surveys on family income, poverty, and eligibility estimates. *Health Services Research*, 40, 1534–1552. 10.1111/j.1475-6773.2005.00416.x [PubMed: 16174146]
- Deaton A (2011). The financial crisis and the well-being of America. In *Investigations in the Economics of Aging Volume* (pp. 343–368).
- Deaton A, & Stone AA (2016). Understanding context effects for a measure of life evaluation: How responses matter. *Oxford Economic Papers*, 68, 861–870. 10.1093/oep/gpw022 [PubMed: 29129941]
- Dolan P, & White MP (2007). How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science*, 2, 71–85. 10.1111/j.1745-6916.2007.00030.x [PubMed: 26151920]
- Egidi G, & Caramazza A (2014). Mood-dependent integration in discourse comprehension: Happy and sad moods affect consistency processing via different brain networks. *NeuroImage*, 103, 20–32. 10.1016/j.neuroimage.2014.09.008 [PubMed: 25225000]
- Hauser DJ, & Schwarz N (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48, 400–407. 10.3758/s13428-015-0578-z [PubMed: 25761395]
- Huff C, & Tingley D (2015). “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics*, 2, 10.1177/2053168015604648
- Kahneman D, Krueger AB, Schkade DA, Schwarz N, & Stone AA (2004). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science*, 306, 1776–1780. 10.1126/science.1103572 [PubMed: 15576620]
- Keith MG, Tay L, & Harms PD (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, 8, 10.3389/fpsyg.2017.01359
- Krosnick JA, & Alwin DF (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, 201–219. 10.1086/269029
- McCredie MN, & Morey LC (2018). Who are the Turkers? A characterization of MTurk workers using the Personality Assessment Inventory. *Assessment*, 1–8. 10.1177/1073191118760709
- Paolacci G, & Chandler J (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184–188. 10.1177/0963721414531598
- Ramsey SR, Thompson KL, McKenzie M, & Rosenbaum A (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, 58, 354–360. 10.1016/j.chb.2015.12.049
- Schwarz N (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105. 10.1037//0003-066X.54.2.93
- Shank DB (2016). Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk. *American Sociologist*, 47, 47–55. 10.1007/s12108-015-9266-9
- Shapiro DN, Chandler J, & Mueller PA (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, 1, 213–220. 10.1177/2167702612469015
- Smith NA, Sabat IE, Martinez LR, Weaver K, & Xu S (2015). A convenient solution: Using MTurk to sample from hard-to-reach populations. *Industrial and Organizational Psychology*, 8, 220–228.
- Smith SM, Roster CA, Golden LL, & Albaum GS (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69, 3139–3148. 10.1016/j.jbusres.2015.12.002

- Stephoe A, Deaton A, & Stone AA (2015). Subjective wellbeing, health, and ageing. *The Lancet*, 385, 640–648. 10.1016/S0140-6736(13)61489-0
- Stone AA, Schwartz JE, Broderick JE, & Deaton A (2010). A snapshot of the age distribution of psychological well-being in the United States. *Proceedings of the National Academy of Sciences*, 107, 9985–9990. 10.1073/pnas.1003744107
- Thorley C, Dewhurst SA, Abel JW, & Knott LM (2016). Eyewitness memory: The impact of a negative mood during encoding and/or retrieval upon recall of a non-emotive event. *Memory*, 24, 838–852. 10.1080/09658211.2015.1058955 [PubMed: 26273852]
- Walters K, Christakis DA, & Wright DR (2018). Are Mechanical Turk worker samples representative of health status and health behaviors in the U.S.? *PLoS ONE*, 13, 1–10. 10.1371/journal.pone.0198835
- [dataset] Whillans A (2017). DATA - Buying time project study data.
- Whillans AV, Dunn EW, Smeets P, Bekkers R, & Norton MI (2017). Buying time promotes happiness. *Proceedings of the National Academy of Sciences*, 114, 8523–8527. 10.1073/pnas.1706541114

Highlights:

Amazon Mechanical Turk (MTurk) samples are frequently used in behavioral research.

MTurk participants have lower subjective well-being (SWB) than other samples.

Selected demographic differences did not explain the lower SWB in MTurk sample.

Further studies are needed to understand lower SWB in MTurk samples.

Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you.

On which step of the ladder would you say you personally feel you stand at this time?

- ☐ 10 Best possible
- ☐ 9
- ☐ 8
- ☐ 7
- ☐ 6
- ☐ 5
- ☐ 4
- ☐ 3
- ☐ 2
- ☐ 1
- ☐ 0 Worst possible

Figure 1.
The Ladder question presented in online surveys.

Figure 2.

Ladder scores by Income (Left Panel) or Race (Right Panel) in the MTurk, UAS Internet panel, and Gallup-Sharecare samples.

Figure 3.

Ladder means from this study (panel A), Whillans et al. (2017) (panel B), and Bussoni & Samani (2018) (panel C). Error bars represent 95% confidence intervals.

Table 1.

Response categories for demographic variables in all samples and in the harmonized version.

Variable	MTurk sample	UAS sample	Gallup sample	Harmonized
Education	Up to 8 th grade	Less than 1st grade	Less than high school	Less than high school
	8 th to 11 th grade	Up to 4th grade 5th or 6th grade 7th or 8th grade 9th grade 10th grade 11th grade 12th grade no-diploma		
	High school graduate	High school graduate or GED	High school degree or diploma	High school graduate
	Some college	Some college-no degree Associate college degreeoccupation/ vocation program Associate college degreeacademic program	Technical/Vocation school Some college	Some college
	College graduate Master's degree Doctoral degree	Bachelor's degree Master's degree Professional school degree Doctorate degree	College graduate Post graduate work or degree	College graduate Post graduate
Race	White	White	White	White
	African American	Black	Black	Black
	Native American Asian Pacific Islander Other More than one race	American Indian or Alaska Native Asian Hawaiian/Pacific Islander Mixed	Asian Hispanic Other	Other
Income	< \$20,000	< \$5,000 \$5,000 – \$7,499 \$7,500 – \$9,999 \$10,000 – \$12,499 \$12,500 – \$14,999 \$15,000 – \$19,999	< \$720 \$720 – \$5,999 \$6,000 – \$11,999 \$12,000 – 23,999	< \$20,000
	\$20,000 – \$34,999	\$20,000 – \$24,999 \$25,000 – \$29,999 \$30,000 – \$34,999	\$24,000 – \$35,999	\$20,000 – \$34,999
	\$35,000 – \$49,999	\$35,000 – \$39,999 \$40,000 – \$49,999	\$36,000 – \$47,999	\$35,000 – \$49,999
	\$50,000 – \$74,999	\$50,000 – \$59,999 \$60,000 – \$74,999	\$48,000 – \$59,999 \$60,000 – \$89,999	\$50,000 – \$74,999
	> \$75,000	\$75,000 – \$99,999 \$100,000 – \$149,999 > \$150,000	\$90,000 – \$119,999 > \$120,000	> \$75,000
Marital Status	Never married	Never married	Single/Never married	Never married
	Married	Married (spouse lives with me) Married (spouse lives elsewhere)	Married	Married
	Living with partner	Living with partner	Domestic partnership/Living with partner (not legally married)	Living with partner
	Separated	Separated	Separated	Separated
	Divorced	Divorced	Divorced	Divorced
	Widowed	Widowed	Widowed	Widowed

Table 2.

Ladder and demographic variables by sample.

	MTurk (n=933)	UAS Panel (n=221)	Gallup-Sharecare (n=137,185)
Ladder (M; SD)	5.26 (2.00)	6.92 (1.94)	7.11 (1.89)
Age (M; SD)	37.6 (12.1)	49.2 (14.9)	51.9 (16.4)
20–29	29%	11%	13%
30–39	35%	22%	14%
40–49	17%	14%	15%
50–59	12%	20%	20%
60–69	5%	24%	23%
70–80	2%	8%	16%
Gender (%)			
Female	56%	64%	50%
Male	44%	36%	50%
Education (%)			
< High School Graduate	1%	6%	5%
High School Graduate	11%	16%	21%
Some College	32%	40%	30%
College Graduate	42%	22%	24%
Post Graduate	14%	16%	20%
Race (%)			
White	81%	82%	77%
Black	8%	8%	9%
Other	12%	10%	14%
Marital Status (%)			
Never Married	34%	13%	19%
Married	44%	60%	53%
Living with Partner	12%	10%	5%
Separated	1%	1%	2%
Divorced	8%	13%	11%
Widowed	1%	3%	10%
Income (%)			
< \$20,000	13%	18%	16%
\$20,000 - \$34,999	21%	12%	12%
\$35,000 - \$49,999	18%	11%	9%
\$50,000 - \$74,999	24%	23%	29%
\$75,000	26%	36%	35%

Table 3.

ANOVA results for Ladder differences between samples before (Model 1) and after (Model 2) entering demographic covariates

	Model 1	Model 2
Sample		
MTurk	ref	ref
UAS Internet Panel	1.66	1.50
Gallup	1.86	1.65
Age		
20–29		ref
30–39		–.14
40–49		–.23
50–59		–.19
60–69		.13
70–80		.44
Gender		
Male		ref
Female		.26
Education		
< High School Graduate		ref
High School Graduate		–.04
Some College		–.06
College Graduate		.15
Post Graduate		.28
Race		
White		ref
Black		.16
Other		.28
Marital Status		
Never Married		ref
Married		.33
Living with Partner		.11
Separated		–.24
Divorced		–.06
Widowed		–.07
Income		
< \$20,000		ref
\$20,000 - \$34,999		.34
\$35,000 - \$49,999		.57

	Model 1	Model 2
Sample		
\$50,000 - \$74,999		.89
\$75,000		1.32
Constant		4.22
F-test	F(2, 138,336)= 447.6***	F(23, 138,314)= 788.7***
Variance explained	0.6%	11.5%

Note. Only participants with complete data on all variables included in both models. UAS = Understanding America Study