

Amazon's Mechanical Turk:

A New Source of Inexpensive, Yet High-Quality, Data?

Michael D. Buhrmester, Tracy Kwang, and Samuel D. Gosling

The University of Texas at Austin

Perspectives on Psychological Science, in press

This is an unedited manuscript that has been accepted for publication. It will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form.

Abstract Word Count: 140

Ms + References Word Count: 1369

Abstract

Amazon's Mechanical Turk (MTurk) is a relatively new website that contains the major elements required to conduct research: an integrated participant compensation system, a large participant pool, and a streamlined process of study design, participant recruitment, and data collection. Here, we describe and evaluate the potential contributions of MTurk to psychology and other social sciences. Findings indicate that: (a) MTurk participants are slightly more representative of the U.S. population than are standard Internet samples and are significantly more diverse than typical American college samples; (b) participation is affected by compensation rate and task length but participants can still be recruited rapidly and inexpensively; (c) realistic compensation rates do not affect data quality; and (d) the data obtained are at least as reliable as those obtained via traditional methods. Overall, MTurk can be used to obtain high-quality data inexpensively and rapidly.

Keywords: Amazon Mechanical Turk, Internet, online, web, data collection, research methods

Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?

Amazon's Mechanical Turk (www.MTurk.com) is a novel, open marketplace for getting work done online. Here, we describe and evaluate the potential contributions that MTurk may make in psychology and other social sciences as a vehicle for Web-based data-collection.

I. Introduction to MTurk

Q1.1 How Does MTurk Work?

MTurk functions as a one-stop shop for getting work done, bringing together the people and tools that enable task creation, labor recruitment, compensation, and data collection. The site boasts a large, diverse workforce consisting of over 100,000 users from over 100 countries who complete tens of thousands of tasks daily (Pontin, 2007). Individuals register as “requesters” (task creators) or “workers” (paid task completers). Requesters can create and post virtually any task that can be done at a computer (i.e., surveys, experiments, writing, etc.) using simple templates, technical scripts, or linking workers to external online survey tools (e.g., SurveyMonkey). Workers can browse available tasks and are paid upon successful completion of each task. Requesters can refuse payment for subpar work. Being refused payment has negative consequences for workers because requesters can limit their tasks to workers with low refusal rates.

Q1.2 How are Workers Compensated?

Requesters deposit money into an account using a credit card. Requesters set the compensation amount prior to posting a task; payments can be awarded automatically or manually based on the quality of each worker submission. Amazon charges a 10% commission.

Q1.3 Why do Workers Participate?

Compensation in MTurk is monetary, but the amount awarded is typically small (e.g., nickels and dimes for 5-10 minute tasks). Our analyses (see online supporting materials) of worker motivation suggest that they are internally motivated (e.g., for enjoyment).

II. Evaluating the Quality of MTurk Data

Q2.1 How Do MTurk Samples Compare to Other Samples?

Commentators have long lamented the heavy reliance on American college samples in the field of psychology (Sears, 1986) and more generally those from a small sector of humanity (Henrich et al., in press). Recent evidence suggests that collecting data via the Internet, although far from perfect, can reduce the biases found in traditional samples (Gosling et al., 2004).

To examine how MTurk samples compare to the diversity of standard Internet samples, we compared the demographics of 3,006 MTurk participants with those in a large Internet sample (Gosling et al., 2004). MTurk participants came from over fifty different countries and all fifty U.S. states. Gender splits were similar in the standard Internet (57% female) and MTurk (55% female) samples. A greater percentage of MTurk participants were non-white (36%) and almost equally non-American (31%) compared to the Internet sample (23% and 30%, respectively). MTurk participants were older ($M = 32.8$; $SD = 11.5$) than the Internet participants ($M = 24.3$; $SD = 10.0$). In short, MTurk participants were slightly more representative of the U.S. population than are standard Internet samples and significantly more diverse than typical American college samples.

Q2.2 How Do Compensation Amount and Task Length Affect Participation Rates?

MTurk's major appeal is its potential for collecting data inexpensively and rapidly. To investigate participant response rates at various compensation levels and task lengths, and to explore the tradeoffs between these parameters, we administered personality questionnaires via

MTurk in a 3 X 3 design, crossing compensation level (2, 10, or 50 cents) with estimated task-completion time (5, 10, and 30 minutes).

There was a main effect of compensation level, $F(2,6) = 20.67, p < .01$, with participation rates lowest in the 2 cent payment (see Table 1). With the exception of the 2 cent condition (due to a possible floor effect), there was a main effect of survey length such that response rates were lowest for the 30 minute survey, $F(1,6) = 7.05, p < .05$. Note that although participation rates decreased as a function of both payment amount and survey length, we were still able to recruit participants for all conditions.

To explore the lower limits of compensation amount for task completion, we tested whether MTurk workers would complete a task for the lowest allowable payment rate – *a penny*. We posted a task that paid workers one cent for answering two pieces of information: age and gender. In 33 hours, we collected 500 responses, or about 15 participants per hour. These results demonstrate that workers are willing to complete simple tasks for virtually no compensation, again suggesting that workers are not driven primarily by financial incentives.

These analyses suggest that participants can be recruited rapidly and inexpensively. Participation rates are sensitive to compensation amounts and time commitments, but our findings demonstrate that it is possible to collect decent-sized samples via MTurk for mere dollars. Even when offering just two cents for a 30 minute task, we accumulated 25 participants, albeit at a slower rate (i.e., in about five hours of posting time). Moreover, by increasing the compensation just slightly (e.g., to 50 cents) we were able to obtain the same number of participants in less than two hours of posting time.

Q2.3 How Does Compensation Amount Affect Data Quality?

To examine compensation level effects on data quality, we computed alpha reliabilities for data collected at three levels of compensation (2, 10, and 50 cents) in a set of six personality questionnaires administered to MTurk participants (e.g., attachment styles and Big Five personality traits). The mean alphas were within one hundredth of a point across the three compensation levels (see online supporting materials), suggesting that even at low compensation rates, payment levels do not appear to affect data quality; the only drawback appears to be data collection speed (as shown in Q2.2), a finding consistent with previous research on non-survey tasks (Mason & Watts, 2009).

Q2.4: Do MTurk Data Meet Acceptable Psychometric Standards?

The absolute levels of the mean alphas were in the good to excellent range (ranging from $r = .73-.93$; mean $r = .87$ across all scales and compensation levels). Moreover, with three exceptions, the MTurk alphas were within two hundredths of a point of the traditional sample alphas (see online supporting materials). To provide another index of data quality, we estimated test-retest reliabilities in a set of individual difference measures administered three weeks apart via MTurk. Participants were paid twenty cents for completing Wave 1 and fifty cents for Wave 2 (60% completed it). Test-retest reliabilities were very high (ranging from $r = .80$ to $.94$; mean $r = .88$) and compared favorably to test-retest correlations of traditional methods (see online supporting materials).

III. Summary and Conclusions

Our investigation into MTurk as a potential mechanism for conducting research in psychology and other social sciences yielded generally promising findings. The site has the necessary elements to successfully complete a research project from start to finish. Our analyses of demographic characteristics suggest that MTurk participants are at least as diverse and more

representative of non-college populations than those of typical Internet and traditional samples. Most importantly, we found that the quality of data provided by MTurk met or exceeded the psychometric standards associated with published research.

Still, it should be borne in mind that the process of validating MTurk for use by researchers has only just begun. Some of MTurk's current strengths – the open market design and large, diverse participant pool – may change in the future (see online supporting materials for further discussion). That said, if future data continue to be as promising as they have proven here and elsewhere (e.g., Mason & Watts, 2009), we anticipate that MTurk will soon become a major tool for research in psychology and elsewhere in the social sciences.

References

- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*, 93-104.
- Henrich, J., Heine, S.J., & Norenzayan, A. (in press). The weirdest people in the world? *Behavioral and Brain Sciences*.
- Mason, W.A., & Watts, D.J. (2009). Financial incentives and the ‘performance of crowds.’ Proceedings of the Human Computation Workshop. Paris: ACM, June 28, 2009.
- Pontin, J. (2007, March 25). Artificial intelligence: With help from the humans. *The New York Times*. Retrieved from <http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html>.
- Sears, D. O. (1986). College sophomores in the lab: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*, 515–530.

Author Note

Michael Buhrmester, Tracy Kwang, and Sam Gosling, Department of Psychology, The University of Texas at Austin. We thank Matthew L. Brooks and William B. Swann, Jr. for feedback on an earlier version of this manuscript. Correspondence concerning this article should be addressed to any of the authors at the Department of Psychology A8000, 1 University Station, The University of Texas, Austin, Texas, 78712. Electronic mail may be sent to buhrmester@gmail.com

Table 1.

Effects of Compensation Amount and Task Length on Participation Rates (submitted surveys per hour of posting time)

Compensation Amount	Short Survey (5 minutes)	Medium Survey (10 minutes)	Long Survey (30 minutes)
2 cents	5.6	5.6	5.3
10 cents	25.0	14.3	6.3
50 cents	40.5	31.6	16.7

Note. Surveys consisted of a series of demographic questions and personality scales. For the medium length survey, 60 participants were recruited per compensation amount. For the short and long surveys, 25 participants were recruited per compensation amount.

Supplemental materials for

Amazon's Mechanical Turk:

A New Source of Inexpensive, Yet High-Quality, Data?

Michael D. Buhrmester, Tracy Kwang, and Samuel D. Gosling

The University of Texas at Austin

Perspectives on Psychological Science, in press

This is an unedited online supplement for the above manuscript that has been accepted for publication. It will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form.

I. Introduction to MTurk

The name “Mechanical Turk” originates from an 18th century chess-playing “machine” that was showcased around Europe as being able to defeat human opponents. In reality, the machine was a hoax—a human chess master hid inside the machine. Amazon’s 21st century incarnation of the Mechanical Turk functions similarly in spirit, except instead of having a person hidden inside a machine, the human intelligence is on the other end of an Internet connection. MTurk is open to anyone on the Web who knows about the site and is interested in either creating new tasks to be completed by others, or working on other people’s posted tasks.

Q1.2 How are Workers Compensated?

The costs associated with conducting research on MTurk are quite reasonable. Most HITs pay workers fractions of a dollar so an investment of just \$10 would be enough to evaluate the viability of doing a study using the service. Furthermore, Amazon charges a 10% fee on top of

what the workers are paid (rather than charging by the month, as do many commercial data-collection and survey sites). Such an approach allows a curious researcher to undertake some basic exploratory research without having to make a significant financial commitment.

Q1.3 Why do Workers Participate?

The modest compensation rates raise the question of whether MTurk workers are driven primarily by financial incentives or whether other motivations play a role. To gain some insight into worker motivation, we surveyed 187 MTurk workers about the reasons they undertook MTurk HITs (see Table S1). Participants rated the extent to which they agreed with each statement on a scale ranging from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*).

The survey findings suggest that MTurk workers use the site because they find tasks enjoyable. Note that financial motivations were rated below the scale mean, suggesting that MTurk workers are motivated internally rather than by external rewards; this finding indicates that workers may take the tasks seriously, despite – or perhaps because of – being paid relatively trivial amounts of money.

II. Evaluating the Quality of MTurk Data

Q2.2 How Do Compensation Amount and Task Length Affect Participation Rates?

Table S2 (Table 1 in the original manuscript) shows the participation rates (number of participants recruited per hour of posting time) across three levels of compensation and estimated task-completion time. The time of day and time of week when the questionnaires were administered were counterbalanced across conditions.

Q2.3 How Does Compensation Amount Affect Data Quality?

Measures of self-concept clarity (Campbell et al., 1996), attachment styles (AAQ; Simpson, Rholes, & Nelligan, 1992), self-liking and self-competence (Tafarodi & Swann, 2001),

global self-esteem (RSES; Rosenberg, 1965), Social Dominance Orientation (SDO; Pratto, Sidanius, Stallworth, & Malle, 1994), and Big-Five personality traits (BFI; John & Srivastava, 1999) were administered to MTurk participants ($N = 160$ for most scales). As shown in Table S3, the mean alphas were within one hundredth of a point across the three compensation levels; at the level of the 11 individual scales, the range across compensation levels averaged four hundredths of a point, with a maximum of nine hundredths of a point.

Q2.4: Do MTurk Data Meet Acceptable Psychometric Standards?

To measure test-retest reliabilities, we obtained self-reports from 116 MTurk participants on measures of political conservatism and liberalness, global self-esteem (RSES; Rosenberg, 1965), Social Dominance Orientation (SDO; Pratto, Sidanius, Stallworth, & Malle, 1994), and Big-Five personality traits (BFI; John & Srivastava, 1999). Participants who agreed to participate in a follow-up survey were contacted three weeks later to complete Wave 2 of the same scales. Participants were paid twenty cents for completing the Wave 1 survey, and fifty cents for completing Wave 2. 60% of respondents ($N=70$) completed the second wave. Results, which are presented in Table S4, show that the test-retest reliabilities were very high, ranging from $r = .80$ to $.94$, with a mean of $.88$. These reliabilities compare favorably to reports of test-retest correlations obtained using traditional methods.

III. The Future of MTurk-Based Research in Psychology and Other Social Sciences

In this section, we discuss the types of research designs that might be successfully conducted using MTurk. We also consider current limitations of using MTurk and present some suggestions for addressing them.

Q3.1 What types of research can be done on MTurk?

A broad range of research paradigms could be adapted for use on MTurk, although some more easily than others. Any paradigm currently on the Internet can utilize the recruitment and payment system of MTurk to direct MTurk workers to researchers' existing projects. For example, if a researcher is having difficulty attracting sufficient traffic online, a simple post could be created on MTurk instructing participants to navigate to the study. At the end of the study, researchers can provide participants with a unique completion code (or have them create one themselves), and instruct them enter the code in a textbox on the MTurk site before they submit the HIT as complete. This completion code strategy allows one to be able to link participant data to their anonymized MTurk worker ID, a necessity to be able to reject workers who provide significantly incomplete data. This strategy may also deter workers from submitting the HIT as complete without having actually participated in the off-site study.

Other research designs that are common to psychological science can be conducted on MTurk. Researchers exploring lay understandings of a construct can ask for open-ended responses and even use other MTurk workers to code them. Researchers who need to construct a new scale and validate it in a separate sample can conduct both steps on MTurk. Cross-sectional, experimental, and even longitudinal studies (with acceptable follow-up response rates) can be conducted on MTurk as well. With slightly more technical knowledge, tasks involving the coding of audio and visual data are possible. Time-consuming projects such as locating sources of data or long transcriptions can be spread out over a large number of MTurk workers and completed in much less time than is possible with traditional approaches that rely on small teams of dedicated (but bored) research assistants.

Q3.2 What are MTurk's current limitations and how can they be addressed?

Despite MTurk's promise for research in psychology, it is important to consider its limitations too and, where possible, devise ways to address them.

3.2.1: Lack of control. As with most Internet-based methods, MTurk studies can exert only minimal control over participants' environments compared to lab studies. However, this lack of control may be a double-edged sword: The quality of data may suffer to an unknown extent due to the absence of standardized, controlled testing conditions. For some topics, however, demand characteristics, self-presentation biases, and experimenter biases will be lower in MTurk studies than in traditional lab studies. Moreover, recent research on the consistency of findings obtained across traditional lab and Internet-based studies suggests that the discrepancies between the two methods are not as great as is widely imagined (Berrens et al., 2003; Buchanan, Johnson, & Goldberg, 2005; Coles, Cook, & Blake, 2007; Myerson & Tryon, 2004).

3.2.2: Not representative of populations. Some of MTurk's apparent advantages over other Internet methods could prove to be a source of concern. As noted above, MTurk participants tend to be slightly more diverse than other Internet samples and significantly more diverse than traditional samples. However, MTurk participants are not representative of the American population, or any other population for that matter. More than a third of the MTurk participants in our research reported being from a country other than the U.S. so it may be tempting to conduct cross-cultural research using MTurk. However, we caution against such research in non-English speaking countries because MTurk is currently based only in English, and English speakers from non-English speaking countries are unlikely to comprise representative samples.

3.2.3: Deceptive responding. The total anonymity afforded by online studies has been identified by some commentators as a disadvantage compared to traditional methods (Skitka &

Sargis, 2005). The commentators' major concern is that participants may simply lie about themselves in a way that would be difficult to detect. It is quite possible that some MTurk participants are being dishonest in their responses, but this may be reduced through a number of MTurk features. Unlike many other web-based projects, MTurk participants can complete a given HIT only once without going to the lengths of creating additional MTurk accounts. Moreover, MTurk's "Qualifications" feature allows requesters to select only the workers that meet criteria specified by the researcher. The number of built-in selection qualifications offered by MTurk are quite limited (country and approval rating) but requesters can add their own qualifications using MTurk's more advanced scripting toolbox features. Even the two default qualification options can help reduce lying in some cases. For example, if a researcher wants to sample only Americans, a qualification can be set to allow only workers who initially said they were from the U.S. when originally creating their worker account. This information is not something the workers can change to match the qualifications requirements of each HIT.

3.2.4: Limits to types of possible research. There are some types of research designs that are difficult or impossible to conduct online. Physiological measurements or any sort of physical manipulations would be impossible. Studies involving live person-to-person interactions, however, may become easier as person-to-person audio and video streaming technology becomes more widely adopted.

3.2.5: Market shifts. Finally, MTurk may have a unique limitation. As an open marketplace, the basic economic forces of supply and demand are at play. Shifts in these dynamics could result in shifts in some of the basic parameters that we have evaluated (study cost, participation rates, sample characteristics, and data quality). For instance, workers unsatisfied with the low levels of compensation may start to leave the site.

Other shifts such as the demographics of workers are quite possible and will require close monitoring by researchers. Also, MTurk is controlled by Amazon.com and not researchers, so the future of the site is entirely out of researchers' hands. If Amazon.com decides to pull the plug, researchers will have to look elsewhere or attempt to emulate the platform from scratch.

References

- Berrens, R.P., Bohara, A.K., Jenkins-Smith, H., Silva, C., & Weimer, D.L. (2003). The advent of Internet surveys for political research: A comparison of telephone and internet samples. *Political Analysis, 11*, 1-22.
- Bosson, J., Swann, W. B., Jr., & Pennebaker, J. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631-643.
- Buchanan, T. Johnson, J.A., Goldberg, L.R. (2005). Implementing a Five-Factor Personality Inventory for Use on the Internet. *European Journal of Psychological Assessment, 21*(2), 115-127.
- Campbell, J. D., Trapnell, P. D., Heine, S. J., Katz, I. M., Lavalley, L. F., & Lehman, D. R. (1996). Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology, 70*(1), 141-156.
- Coles, M.E., Cook, L.M., & Blake, T.R. (2007). Assessing obsessive compulsive symptoms and cognitions on the internet: Evidence for the comparability of paper and Internet administration. *Behavior Research and Therapy, 45*, 2232-2240.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*, 504-528.
- John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102-138). New York: Guilford Press.

- Meyerson, P., & Tryon, W.W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, 35(4), 614-620.
- Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, 67, 741-763.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Simpson, J. A., Rholes, W. S., & Nelligan, J. S. (1992). Support seeking and support giving within couples in an anxiety-provoking situation: The role of attachment styles. *Journal of Personality and Social Psychology*, 62(3), 434-446.
- Skitka, L.J., & Sargis, E.G. (2006). The internet as a psychological laboratory. *Annual Review of Psychology*, 57, 529–555.
- Tafarodi, R. W., & Swann, W. B., Jr. (2001). Two-dimensional self-esteem: Theory and measurement. *Personality and Individual Differences*, 31, 653-673.

Table S1

Self-Reported Motives for MTurk Workers

Question	<i>M</i>	<i>SD</i>
Why do you use MTurk?		
1. Enjoy doing interesting tasks	5.08	1.56
2. To kill time	4.85	1.88
3. To have fun	4.43	1.68
4. To make money	3.39	1.94
5. To gain self-knowledge	3.13	1.88

Note. N = 187. Measured using a 1 (*Strongly Disagree*) to 7 (*Strongly Agree*) point Likert Scale.

Table S2 (Table 1 in original ms)

Participation Rates (submitted surveys per hour of survey posting time) as a Function of Compensation Amount and Task Length

Compensation Amount	Short Survey (5 minutes)	Medium Survey (10 minutes)	Long Survey (30 minutes)
2 cents	5.6	5.6	5.3
10 cents	25.0	14.3	6.3
50 cents	40.5	31.6	16.7

Note. Surveys consisted of a series of demographic questions and personality scales. For the medium length survey, 60 participants were recruited per compensation amount. For the short and long surveys, 25 participants were recruited per compensation amount.

Table S3

Reliability Alphas Between Samples

Scale	<u>MTurk</u>				<u>Standard Internet</u>
	2 cents	10 cents	50 cents	Average	
SDO	.93	.89	.93	.92	.91
RSES	.90	.90	.91	.90	.91
BFI Extraversion	.86	.88	.85	.86	.87
BFI Agreeableness	.76	.73	.82	.77	.77
BFI Conscientiousness	.86	.86	.82	.85	.77
BFI Emotional Stability	.89	.89	.87	.88	.85
BFI Openness	.80	.90	.80	.83	.79
Clarity	.87	.92	.93	.91	.90
Avoidant	.81	.85	.81	.82	.84
Anxious	.85	.81	.81	.82	.81
SLCS	.93	.92	.93	.93	.92
Mean	.87	.88	.87	.87	.86

Note. For MTurk data, N = 160 for all scales except N = 74 for Clarity, Avoidant, Anxious, and SLCS. All MTurk data were collected over a two-week period from January through February 2010. BFI = Big Five Inventory, SDO = Social Dominance Orientation, RSES = Rosenberg Self-Esteem Scale, SLCS = Self-Liking and Competence Scale, AAQ= Adult Attachment Questionnaire, Comparison samples for the BFI, SDO, and RSES came from a large sample of college undergraduates (N = 1822) collected by Gosling, Rentfrow, and Swann (2003). Comparison samples for self-concept clarity and AAQ consisted of 116 participants collected online by the second author through posting ads on Yahoo groups, Facebook, and Craigslist during February 2009.

Table S4

Test-Retest Reliability Correlations in MTurk Data and Traditional Data Samples

Variable	MTurk	Previous literature
Politically Conservative	.85**	n/a
Politically Liberal	.80**	n/a
BFI Extraversion	.94**	.82**
BFI Agreeableness	.87**	.76**
BFI Conscientiousness	.86**	.76**
BFI Emotional Stability	.92**	.83**
BFI Openness	.90**	.80**
RSES	.87**	.80**
SDO	.87**	.81**

Note. ** $p < .01$. Out of the original 116 Wave 1 MTurk participants, 70 completed Wave 2 (60%); BFI = Big Five Inventory. The variables “politically conservative” and politically liberal” were single item scales, and comparable published data was not available (n/a). The comparison sample for the BFI scales was comprised of 114 undergraduates from Gosling et al. (2003). The comparison sample for the RSES was comprised of 84 undergraduates from Bosson, Swann, & Pennebaker (2000). The comparison sample for the SDO scale was comprised of 25 undergraduates from Pratto et al. (1994).