Online experiments workshop
(EPC 2015)

Drew Hendrickson
Dan Navarro
Steve Langsford
(Amy Perfors)
Why are we doing this?
Google Scholar results for “Mechanical Turk”
Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling
Department of Psychology, University of Texas at Austin

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. In this article, we describe and evaluate the potential contributions of MTurk to psychology and other social sciences. Findings indicate that (a) MTurk participants are slightly more demographically diverse 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
Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research

Matthew J. C. Crump\textsuperscript{1*}, John V. McDonnell\textsuperscript{2}, Todd M. Gureckis\textsuperscript{2}

\textsuperscript{1} Department of Psychology, Brooklyn College of CUNY, Brooklyn, New York, United States of America, \textsuperscript{2} Department of Psychology, New York University, New York, New York, United States of America

Abstract

Amazon Mechanical Turk (AMT) is an online crowdsourcing service where anonymous online workers complete web-based tasks for small sums of money. The service has attracted attention from experimental psychologists interested in gathering human subject data more efficiently. However, relative to traditional laboratory studies, many aspects of the testing environment are not under the experimenter’s control. In this paper, we attempt to empirically evaluate the fidelity of the AMT system for use in cognitive behavioral experiments. These types of experiment differ from simple surveys in that they require multiple trials, sustained attention from participants, comprehension of complex instructions, and millisecond accuracy for response recording and stimulus presentation. We replicate a diverse body of tasks from experimental psychology including the Stroop, Switching, Flanker, Simon, Posner Cuing, attentional blink, subliminal priming, and category learning tasks using participants recruited using AMT. While most of replications were qualitatively successful and validated the approach of collecting data anonymously online using a web-browser, others revealed disparity between laboratory results and online results. A number of important lessons were encountered in the process of conducting these replications that should be of value to other researchers.


Editor: Sam Gilbert, University College London, United Kingdom

Received November 20, 2012; Accepted January 21, 2013; Published March 13, 2013

Copyright: © 2013 Crump et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mcrump@brooklyn.cuny.edu
Should We Trust Web-Based Studies?

A Comparative Analysis of Six Preconceptions About Internet Questionnaires

Samuel D. Gosling and Simine Vazire  
Sanjay Srivastava  
Oliver P. John  
University of Texas at Austin  
Stanford University  
University of California, Berkeley

The rapid growth of the Internet provides a wealth of new research opportunities for psychologists. Internet data collection methods, with a focus on self-report questionnaires from self-selected samples, are evaluated and compared with traditional paper-and-pencil methods. Six preconceptions about Internet samples and data quality are evaluated by comparing a new large Internet sample (N = 361,703) with a set of 510 published traditional samples. Internet samples are shown to be relatively diverse with respect to gender, socioeconomic status, geographic region, and age. Moreover, Internet findings generalize across presentation formats, are not adversely affected by nonserious or repeat responders, and are consistent with findings from traditional methods. It is concluded that Internet methods can contribute to many areas of psychology.

those obtained with most traditional techniques. In addition, Internet methods offer a variety of more mundane, but practically significant, benefits such as dispensing with the need for data entry and being relatively inexpensive. However, these benefits cannot be realized until researchers have first evaluated whether this new technique compromises the quality of the data.

Although many researchers have begun using this new tool, its benefits and potential obstacles have gone largely unexamined. Previous researchers have addressed technical issues—the “how to” of Internet data collection (Birnbaum, 2001; Dillman, 1999; Dillman, Tortora, & Bowker, 1998; Fraley, 2004; Kieley, 1996; Morrow & McKee, 1998)—or have speculated about the pros and cons (Hewson, Laurent, & Vogel, 1996; Kraut et al., 2003; Michalak & Szabo, 1998; Schmidt, 1997). However, few have tested empiri-
Collecting response times using Amazon Mechanical Turk and Adobe Flash

Travis Simcox • Julie A. Fiez

Published online: 14 May 2013
© Psychonomic Society, Inc. 2013

Abstract Crowdsourcing systems like Amazon’s Mechanical Turk (AMT) allow data to be collected from a large sample of people in a short amount of time. This use has garnered considerable interest from behavioral scientists. So far, most experiments conducted on AMT have focused on survey-type instruments because of difficulties inherent in running many experimental paradigms over the Internet. This study investigated the viability of presenting stimuli and collecting response times using Adobe Flash to run ActionScript 3 code in conjunction with AMT. First, the timing properties of Adobe Flash were investigated using a phototransistor and two desktop computers running under several conditions mimicking those that may be present in research using AMT. This experiment revealed some recently, there has been significant interest in conducting behavioral science research on Amazon’s Mechanical Turk (AMT; www.mturk.com). AMT is an online system in which requesters can open accounts and post a wide variety of tasks online, called human intelligence tasks, or HITs. Workers then browse the available HITs, and requesters can pay workers to complete the HITs.

AMT was not created with behavioral science in mind, but it is in many ways suitable for such applications. Mason and Suri (2012) provided a general guide for using AMT in behavioral research and emphasized some benefits of online experimentation using AMT. First, AMT provides a stable pool of participants, in contrast to undergraduate participant
Amazon Mechanical Turk: Gold Mine or Coal Mine?

Karën Fort*  
INIST-CNRS/LIPN

Gilles Adda**  
LIMSI/CNRS

K. Bretonnel Cohen†  
U. Colorado School of Medicine/U. Colorado at Boulder

Recently heard at a tutorial in our field: “It cost me less than one hundred bucks to annotate this using Amazon Mechanical Turk!” Assertions like this are increasingly common, but we believe they should not be stated so proudly; they ignore the ethical consequences of using MTurk (Amazon Mechanical Turk) as a source of labour.

Manually annotating corpora or manually developing any other linguistic resource, such as a set of judgments about system outputs, represents such a high cost that many researchers are looking for alternative solutions to the standard approach. MTurk is becoming a popular one. However, as in any scientific endeavor involving humans, there is an unspoken ethical dimension involved in resource construction and system evaluation, and this is especially true of MTurk.
Using Mechanical Turk to Study Clinical Populations

Danielle N. Shapiro¹, Jesse Chandler², and Pam A. Mueller²
¹Department of Psychology, University of Michigan, Ann Arbor and ²Department of Psychology, Princeton University

Abstract
Although participants with psychiatric symptoms, specific risk factors, or rare demographic characteristics can be difficult to identify and recruit for participation in research, participants with these characteristics are crucial for research in the social, behavioral, and clinical sciences. Online research in general and crowdsourcing software in particular may offer a solution. However, no research to date has examined the utility of crowdsourcing software for conducting research on psychopathology. In the current study, we examined the prevalence of several psychiatric disorders and related problems, as well as the reliability and validity of participant reports on these domains, among users of Amazon’s Mechanical Turk. Findings suggest that crowdsourcing software offers several advantages for clinical research while providing insight into potential problems, such as misrepresentation, that researchers should address when collecting data online.

Keywords
crowdsourcing, Mechanical Turk, clinical methods, Internet research
Why are we doing this?

- We’ve had more experience than I realised…
  - Running online experiments since 2010
  - The lab has probably done 100+ online experiments?
  - Topic varies: categorisation, decision making, language evolution, reaction times, semantics etc
  - Scale varies: 50 to 70000 participants
  - Task length varies: 2 minutes to 1 hour
Why are we doing this?

• We’ve had more experience than I realised…
  • Running online experiments since 2010
  • The lab has probably done 100+ online experiments?
  • Topic varies: categorisation, decision making, language evolution, reaction times, semantics etc
  • Scale varies: 50 to 70000 participants
  • Task length varies: 2 minutes to 1 hour

• And we’ve made some mistakes…
  • It’s a rapidly shifting and unfamiliar area
  • Better to learn from other people’s mistake
What are people asking?

• The big four questions:
  • What kind of studies can I run?
  • Will my data be any good?
  • Is the research ethical?
  • How would I get started?
What are people asking?

- The big four questions:
  - What kind of studies can I run?  Many things, not all
  - Will my data be any good?  Probably, if you’re careful
  - Is the research ethical?  Probably, if you’re careful
  - How would I get started?  We’ll give some pointers
Structure of the workshop

• Introduction (Dan)

• Session 1
  • What can you do with online studies? (Drew)
  • Simple studies with easy tools (Drew)
  • Writing and hosting a web application (Steve)

• Session 2
  • The Amazon Mechanical Turk marketplace (Drew)
  • Research ethics and MTurk studies (Dan)

• Session 3
  • Scientific possibilities from online research (Drew)
  • Assorted tips and tricks for MTurk studies (Dan)