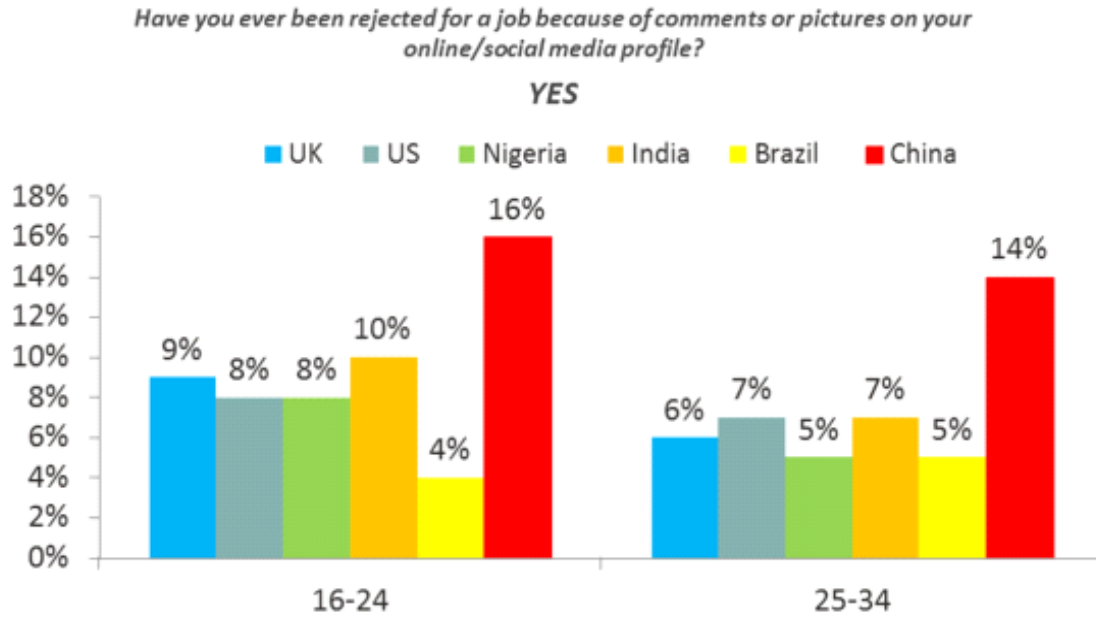


Topic #9

SURVEY RESEARCH

Measurement and assessment of opinions, attitudes, etc. Usually by means of questionnaires and sampling methods.



Source: On Device Research YPCC
 Brazil, China, India, Nigeria, US, UK
 n = 20225 Age 16-34
 May 2013



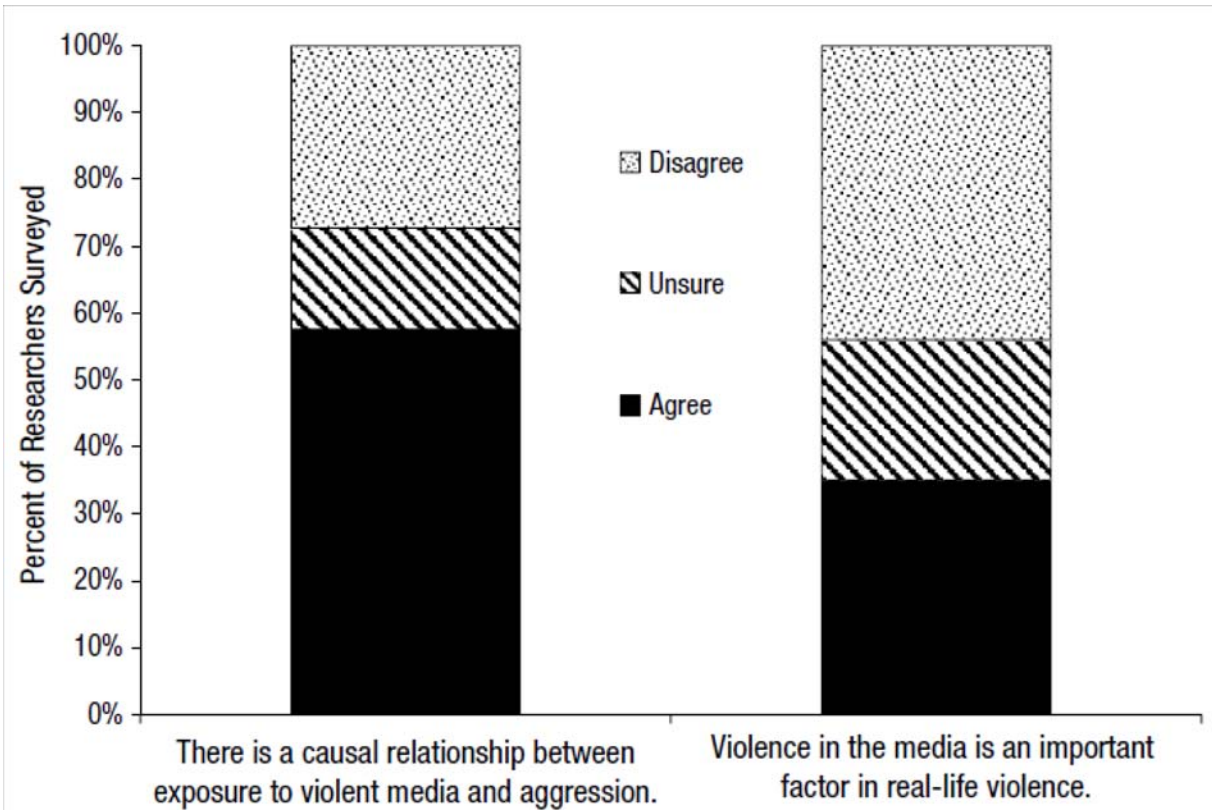


Fig. 1. Media researchers' opinions about the potential negative effects of violent media. Results from a survey of 379 media researchers reported in Bushman, Gollwitzer, and Cruz (in press). Figure was adapted from descriptive data reported in Table 1 of Bushman et al., indicating the number of scientists who agreed, disagreed, or were unsure in their opinions about these questions.

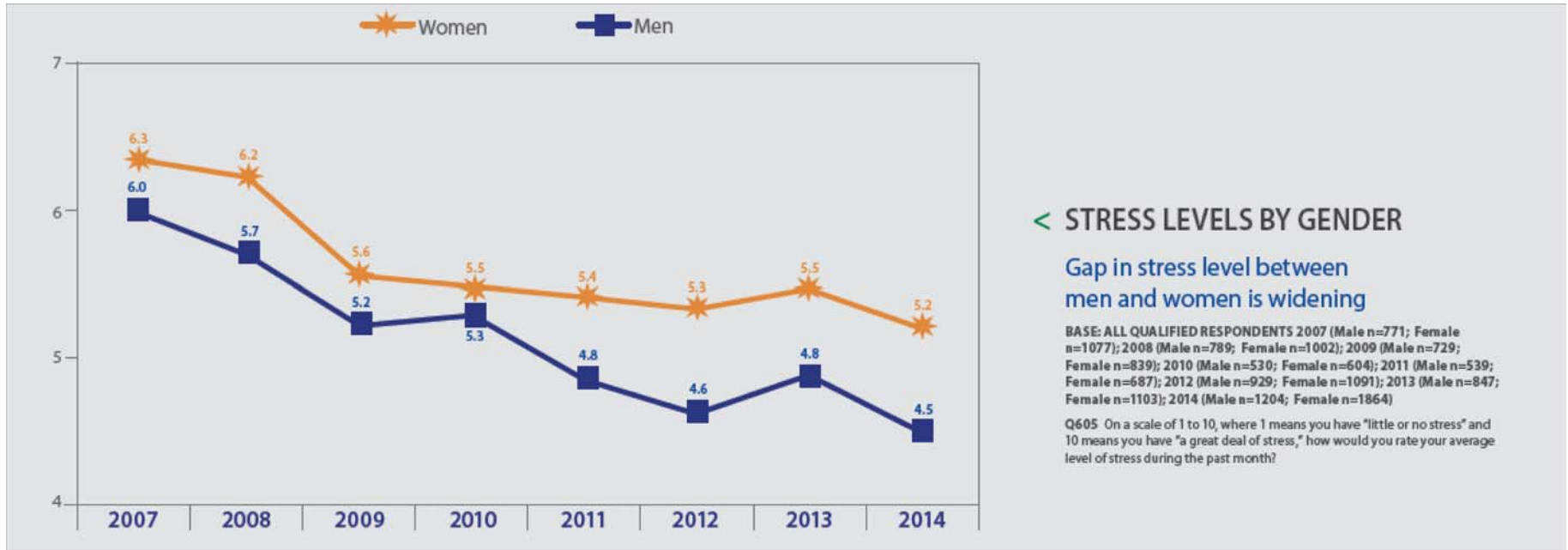
Markey, P. M. (2015). Finding the middle ground in violent video game research: Lessons from Ferguson (2015). *Perspectives on Psychological Science, 10*, 667-670.

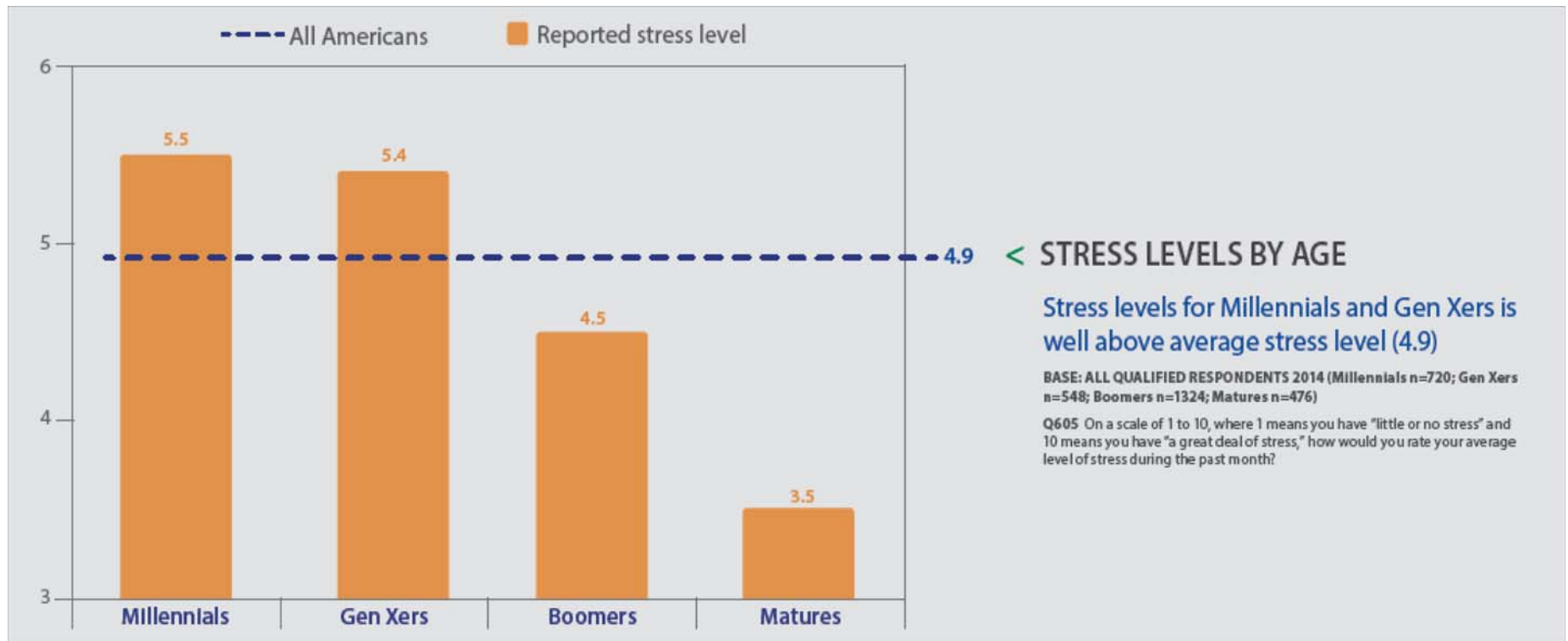
February 4, 2015

American Psychological Association Survey Shows Money Stress Weighing on Americans' Health Nationwide

www.stressinamerica.org







Important Measurement Issues and Potential Problems with Questionnaires and Other Self-Report Measures

1. First determine the purpose of the questionnaire.
 - Ask the target participants for useful information.
 - Anticipate questions of interpretation that may arise.
2. Determine the types of questions.
 - a. **open-ended/constructed-response**—permits the respondents to answer in their own words
 - b. **closed-ended**—limits the respondents to alternatives determined in advance by the designers
3. Item writing
 - Potentially, the questions and items themselves can have a big and major influence on how people will respond.
 - a. Determine the format of the item.
 - construct-response (fill-in or write-in)
 - true/false
 - multiple-choice
 - Likert scales
 - b. Address a single issue per item.

"Are you in favor of increasing student tuition and reducing Transportation Services' budget and allocating this revenue for faculty and staff raises?"

c. Loaded items generate or produce specified responses.

- e.g., item from a mail survey put out by congressman Joe Barton—summer 1988

"A site in Ellis county, in your congressional district, is one of seven national finalists for the superconducting super collider (SCC) project. During this time of budget restraint, do you support programs vital to the future growth of our country such as the SCC?"

Yes = 87%

No = 13%

- During the 1992 presidential election, Ross Perot published a questionnaire in *TV Guide*. His objective was to measure the popularity of the positions he had taken on specified issues. One item read as follows:

Should the president have the Line Item Veto to eliminate waste?

Yes = 97%

Question rewritten and asked of a random sample:

Should the president have the Line Item Veto or not?

Yes = 57%

d. Topic or issue may be "sensitive" which can also have a major influence on how people respond, so avoid bias. Under these conditions, effects of loading are even more pronounced.

- e.g., two items that measure attitude towards abortion:

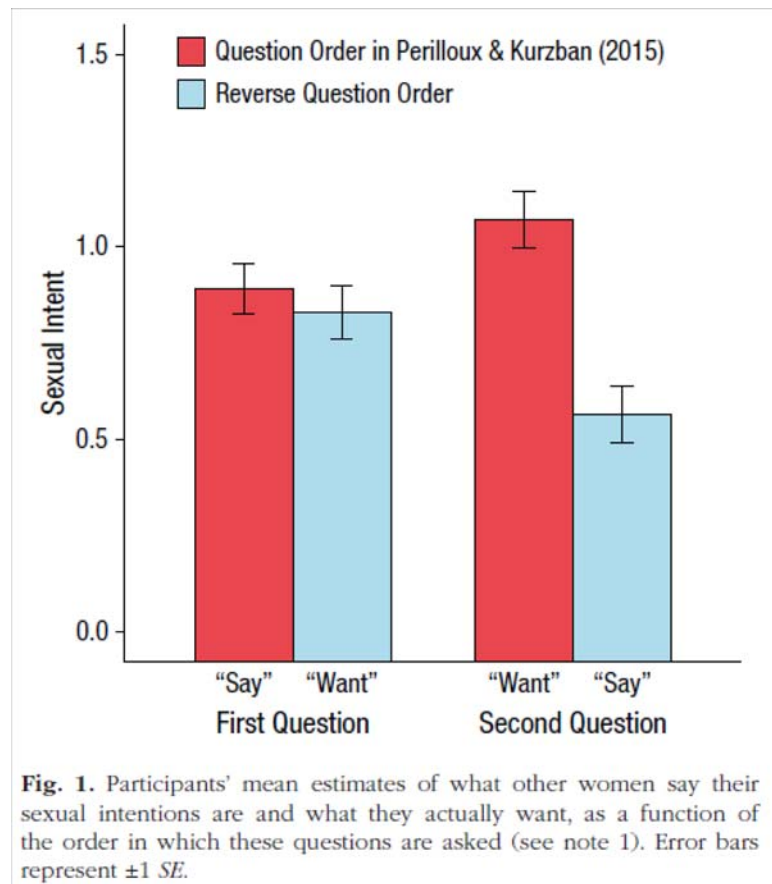
"Do you believe in killing unwanted babies?"

"Should a woman be forced to bear unwanted children?"

e. Effect of question order.

Murray, Murphy, von Hippel, Trivers, & Haselton (2017) PS. A preregistered study of competing predictions suggests that men *do* overestimate women's sexual intent.

Some studies have suggested that men overestimate women's sexual interest. In 2015, Perilloux and Kurzban conducted a study suggesting that this overestimation is in fact an artifact caused by women underestimating their own sexual interest. In this commentary, Murray and colleagues suggest Perilloux and Kurzban's finding may be the product of question order. In the original study, female participants were asked what women say their sexual intentions are ("say" question) before being asked what women's sexual intentions actually are ("want" question"). Murray and colleagues varied the order of these two questions when testing a new set of female participants. When either the "say" or the "want" question was asked first, ratings were equivalent. When participants were asked the "want" question second they adjusted their answer upward and when they were asked the "say" question second they adjusted their answer downward. These findings indicate that the original results were driven by question-order effects.



f. Make the alternatives clear.

- e.g., two items from the Texas Recycles Day Survey used to find out how much students know about recycling on and off campus at Texas A&M.

9. Have you seen recycling bins on campus?

Yes No

10. If you answered yes, how often do you see them?

All the time and everywhere
Eventually
Once upon a time

- Also, do not use negations or percentages.

f. Questions and specified responses to them are not independent—adjacent question effect.

g. There are a variety of ways in which participants' own characteristics may inadvertently alter the research outcome.

1. **Response Styles**—tendencies to respond to questionnaire items in specific ways regardless of content. Biases that are consistent across time and questionnaires.
 - a. **willingness to answer**—some people will not answer items or questions they are unsure about (will leave them blank). Others will go right ahead and guess.
 - *can usually control for this with strong instructions to answer ALL questions*
 - b. **position preference**—when in doubt pick (C)
 - *for knowledge, ability, and other measures with a correct of known "true" score, can control for this by randomization of alternatives*
 - c. **acquiescence or yea- and nay-saying**—tendency to consistently agree or disagree with questionnaire statements or questions regardless of content
 - *controlled for by using method of matched pairs (repeat item and reverse); also controlled by using bi-directional responses*

EXAMPLE (OF ACQUIESCENCE)

Based on the role play instructions, please respond to each question by circling the appropriate number along the line. Please circle the number along the line corresponding to how you think you would feel if you were **John Smith**.

Good	<u>1 2 3 4 5 6 7</u>	Bad
Tense	<u>1 2 3 4 5 6 7</u>	Relaxed
Pleased	<u>1 2 3 4 5 6 7</u>	Displeased
Competent	<u>1 2 3 4 5 6 7</u>	Incompetent
Happy	<u>1 2 3 4 5 6 7</u>	Unhappy

2. **Response Sets**—tendencies to respond to a questionnaire or test content with a particular goal in mind.

The primary example of this is **social desirability**—the most common response set.

- **social desirability**—tendency to present self in a socially desirable manner; tendency to choose specified responses even if they do not represent ones true tendency or opinion.
 - a. **self-deception** occurs when an individual unconsciously views him/herself in an inaccurately favorable light; lack of self-awareness.
 - b. **impression management** refers to a situation in which an individual consciously presents him/herself falsely to create a favorable impression.
- social desirability responding is the tendency to over-report socially desirable personal characteristics and to under-report socially undesirable characteristics.
- also a tendency to present self in test-taking situations in a way that makes self look positive with regard to culturally derived norms and standards.
- e.g., which of the Big Five factors would you expect to be most susceptible impression management/faking effects and why?

conscientiousness

EXAMPLE (OF SOCIAL DESIRABILITY)

Sample items from a test used to select/hire firefighters

5. My friends think I am slightly absent-minded and impractical.
- A. Yes
 - B. Uncertain
 - C. No
10. I prefer a job with _____ opportunity to learn new skills.
- A. a lot of
 - B. some
 - C. little or no

- *control—administer a social desirability responding measure (e.g., MARLOWE-CROWNE, or BALANCED INVENTORY OF DESIRABLE RESPONDING [BIDR] and partial out or drop from sample*

Some tests (e.g., MMPI) have faking scales etc. Anonymous and private collection of data may also help.

h. **Common method variance** and collection of data from single source.

- potential to inflate and confound observed relationships especially where there is a theoretically justifiable reason to expect this.
 - e.g., investigating relationship between task performance, contextual performance, job satisfaction, and organizational commitment and using self-report measures of all variables from employees.

Major Survey Techniques

1. face-to-face interviews
2. telephone interviews
3. mail
4. magazine
5. internet-based surveys & sample recruiting sources
 - online social science research resources
 - Study Response
 - E-Research Global
 - Zoomerang
 - Survey Monkey

Examination of the Equivalence of Self-Report Survey-Based Paper-and-Pencil and Internet Data Collection Methods

Arne Weigold
Notre Dame College

Ingrid K. Weigold and Elizabeth J. Russell
The University of Akron

Self-report survey-based data collection is increasingly carried out using the Internet, as opposed to the traditional paper-and-pencil method. However, previous research on the equivalence of these methods has yielded inconsistent findings. This may be due to methodological and statistical issues present in much of the literature, such as nonequivalent samples in different conditions due to recruitment, participant self-selection to conditions, and data collection procedures, as well as incomplete or inappropriate statistical procedures for examining equivalence. We conducted 2 studies examining the equivalence of paper-and-pencil and Internet data collection that accounted for these issues. In both studies, we used measures of personality, social desirability, and computer self-efficacy, and, in Study 2, we used personal growth initiative to assess quantitative equivalence (i.e., mean equivalence), qualitative equivalence (i.e., internal consistency and intercorrelations), and auxiliary equivalence (i.e., response rates, missing data, completion time, and comfort completing questionnaires using paper-and-pencil and the Internet). Study 1 investigated the effects of completing surveys via paper-and-pencil or the Internet in both traditional (i.e., lab) and natural (i.e., take-home) settings. Results indicated equivalence across conditions, except for auxiliary equivalence aspects of missing data and completion time. Study 2 examined mailed paper-and-pencil and Internet surveys without contact between experimenter and participants. Results indicated equivalence between conditions, except for auxiliary equivalence aspects of response rate for providing an address and completion time. Overall, the findings show that paper-and-pencil and Internet data collection methods are generally equivalent, particularly for quantitative and qualitative equivalence, with nonequivalence only for some aspects of auxiliary equivalence.

Keywords: equivalence testing, quantitative equivalence, qualitative equivalence, auxiliary equivalence, Internet

- Typical overall response rate for survey research is $\approx 30\%$. However, appears to be higher for targeted sampling of paid internet-based samples.
- The quality of the data is a direct function of the response rate.

SAMPLING

- The key to the meaningfulness of any survey is the soundness of the sampling procedure used to generate respondents.
- Examples of inadequate results from poor sampling
 - a. Dewey vs. Truman



Weeks prior to the 1948 election, many leading editorial writers and political columnists relied on early Gallup Polls, which predicted Thomas E. Dewey's win over incumbent Harry S. Truman. Truman's strategy was to bypass the press by taking his case to the people in a "whistlestop" campaign. An issue of the early edition of the Chicago Daily Tribune was handed to Truman after the election. The headline declared "Dewey Defeats Truman." Truman upset Dewey despite having the support of only 15 percent of the nation's daily papers.

- b. Major League Baseball players are selected to the All-Star game by fans at ballparks.



Under the Hood of Mechanical Turk

By K. Andrew DeSoto

When Amazon launched a product called Mechanical Turk (MTurk) just over a decade ago, the e-commerce giant billed it as an online service to enable a marketplace of workers to complete tasks in exchange for payment. But it didn't take long for the product to become a significant research tool in psychological science worldwide.

In 2011, psychological researchers Michael Buhrmester, Tracy Kwang, and APS Fellow Sam Gosling published a paper in *Perspectives on Psychological Science* titled "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" The paper has been cited more than 2,300 times, according to Google Scholar.

And it's easy to see why there is such intense interest in MTurk. Data collection can be much faster online, and MTurk subjects often are compensated at a lower rate than are university students or individuals from other samples, making MTurk research cheaper than average. The service is also an equalizer of sorts — researchers at small schools can have access to large samples that previously were available only at larger research universities. And it's hard to beat the convenience of posting a study online just before bedtime and waking up to a complete data set.

But over the last few years, psychological scientists have begun viewing MTurk with a more critical eye. Many have been concerned about the unique characteristics of the MTurk sample, the appropriate amount to compensate MTurk subjects, and the recent fee increases that Amazon has levied on researchers who use the service.

A Unique Population

A fundamental aspect of MTurk's success is that, in most cases, the subject pool appears to produce quality data. Some of the earliest MTurk research determined that results from online studies often mirror results from lab studies. APS Fellow Jeffrey Karpicke (Purdue University), who uses the tool to study students' learning and memory, believes that this makes the service a valuable tool.

"We are very enthusiastic about MTurk. We have done several experiments both in the lab and on MTurk, and the results look the same," Karpicke said.

Andy DeSoto is the Methodology Fellow at the Association for Psychological Science. He has his PhD in psychology from Washington University in St. Louis and researches metascience, recognition memory, confidence, and collective memory. He can be contacted at apsobserver@psychologicalscience.org or on Twitter at [@kadesoto](https://twitter.com/kadesoto).

"I think the people completing studies on MTurk take the tasks very seriously — probably more than undergraduates doing required experiments for introductory psychology do."


Although results may be consistent across laboratory- and MTurk-based versions of a study, some researchers continue to investigate the idiosyncrasies of the MTurk subject pool.

Scientists who were among the earliest users of MTurk claimed they were capturing samples that were far more representative than the traditional pool of undergraduate students. But now, researchers have determined that MTurk subjects have their own set of distinctive characteristics.

Gabriele Paolacci (Erasmus University Rotterdam, the Netherlands) and Jesse Chandler (University of Michigan and Mathematica Policy Research) summarized these differences in a 2014 article in *Current Directions in Psychological Science*:

"Workers tend to be younger (about 30 years old), over-educated, underemployed, less religious, and more liberal than the general population . . . Within the United States, Asians are overrepresented and Blacks and Hispanics are underrepresented relative to the population as a whole," they wrote. "It should not be treated as representative of the general population."

Many MTurk subjects rely on the service as a source of income. They congregate in online communities, akin to employees meeting around the water cooler. These communities, such as Turkopticon (maintained by researchers at University of California, San Diego), allow MTurk subjects to rate experimenters or labs on a variety of different dimensions, including compensation to subjects. (This author's rating on Turkopticon is found below — it appears that he could improve the amount he pays subjects.) Another digital water cooler can be found in the subreddit "HITs Worth Turking For" on Reddit.com.

AMT Requester Name & ID ▲ ▼	Ratings [] (averaged) ▲ ▼	# of Reports ▲ ▼
Andy DeSoto AIP73UKL5CS8Z HIT Group »	PAY:  2.10 / 5 COMM:  5.00 / 5 FAST:  4.64 / 5 FAIR:  4.42 / 5	24

Whether and how these ratings affect the quality of subjects who participate in a given study is unclear, and these questions should be important topics for future investigations.

And unlike in a laboratory study, in which a researcher can observe a subject going about an experimental task, the convenience of the Internet comes with a level of opaqueness that can

be a challenge to fully grasp. Researchers can use manipulation and attention checks as a way of trying to gauge whether subjects complied with study instructions, but there's no way to guarantee that workers are actually devoting their undivided attention to the study task. In fact, some even admit to completing MTurk tasks while at work.

Another concern with having a core community of workers is that they eventually can become “expert” subjects. Whereas college undergraduates may spend a year or two participating in several studies across many psychology disciplines, MTurk subjects, on average, participate in dozens of studies, sometimes simultaneously or over very short periods of time. A 2015 study by an international team of researchers suggests that these subjects' experiences with common research materials (e.g., the “ball-and-bat problem”) mean that they may not respond as researchers expect them to.

The large number of repeat subjects participating in a given experiment means that the reach of MTurk is narrower than one might expect. A recent study led by psychological scientist Neil Stewart (University of Warwick, England) suggests that researchers using MTurk have an available sample size of 7,300 subjects — greater than the average university research pool but far from the 500,000 workers from 190 countries that Amazon advertises.

Chandler and colleagues followed up with subjects who earlier had completed a series of psychology tasks via MTurk. Subjects were contacted a few days, about a week, or about a month after initial participation and were asked to complete the same tasks online a second time. Some subjects were assigned to the same condition they were in initially, whereas others were assigned to different conditions.

The psychological scientists found that the effects of the experimental manipulations were smaller in the second experiment compared to the first. The decrease in effect size was greatest when subjects were assigned to a different condition in the second study than they were assigned to in the first.

The authors wrote that they found “no direct evidence of a mechanism underlying this decline,” but speculated that several potential factors — including practice effects, cognitive elaboration, and motivation to perform a certain way — could be at work.

Additionally, the large number of repeat subjects participating in a given experiment means that the reach of MTurk is narrower than one might expect. A recent study led by psychological scientist Neil Stewart (University of Warwick, England) suggests that researchers using MTurk have an available sample size of

7,300 subjects — greater than the average university research pool but far from the 500,000 workers from 190 countries that Amazon advertises.

“Amazon Mechanical Turk, and other crowdsourcing platforms, are a great new tool for getting science done,” Stewart said. “But the populations might not be as large as you think; 7,300 workers in your population, often shared with maybe hundreds of other researchers, is great, but it is not that many.”

And many of those research subjects are voicing complaints about the way they are treated and compensated for their participation.

Fees and Compensation

The average rate of pay on MTurk is well below the federal minimum wage of \$7.25 per hour. Earlier this year, *PBS Newshour* produced a piece titled “The Internet's hidden science factory” detailing some of these issues. The program described one worker who estimated completing 20,000 surveys, some presumably psychology experiments, over a 5-year period (that's more than 10 per day).

Pay is low and the tasks are sometimes repetitive, leading some to refer to MTurk as a “digital sweatshop.” That's what led a large group of MTurk workers a year ago to petition Amazon CEO Jeff Bezos to improve worker conditions. One site has even proposed an MTurk code of conduct for academic requesters.

“Many workers consider \$0.10 a minute to be the minimum to be considered ethical,” the document declares. And there may be real repercussions for breaking these guidelines: “Tasks paying less ... are likely to tap into a highly vulnerable work pool and constitutes coercion.” And when accusation of coercion is involved, university Institutional Review Boards may take interest.

“I think there is a very strong case for paying more to hit a living-wage level,” said Stewart.

Money also has become a source of contention for researchers using MTurk. Several months ago, MTurk rolled out a commission increase that's costing researchers more money. A Twitter roundup by the Society for Personality and Social Psychology captures researchers' complaints about the changes.

Prior to this increase, Amazon took 10% of payments to subjects in processing fees. Now, the fee is 40%, assuming the researcher is collecting more than 10 subjects' worth of data at a time. A 200-subject 15-minute study, for instance, would originally have cost \$330, but now costs \$420, assuming a low payment of \$0.10 per minute.

“These changes would be intended to allow us to increase our investment in the marketplace and bring future innovation to Mechanical Turk that will benefit both Requesters and Workers,” stated the official MTurk blog a short time before the price increase occurred. As of yet, no one has noticed any innovations.

Other Options

Researchers unwilling to rely on MTurk alone have the alternative to synthesize MTurk data collection with other kinds of studies. For example, Steven Isley, a quantitative policy analyst at the National Renewable Energy Laboratory in Colorado, uses

MTurk to test ideas before conducting larger-scale studies for the US Department of Energy.

“MTurk has helped us refine many aspects of our research before investing the time and money in a real field trial,” he said, giving an example. “We were going to conduct a field test of a new augmented-reality decision-support tool, and we used MTurk to help us refine our user interface. While the online experiment wasn’t a replacement for the field trial, it helped us understand where the app instructions weren’t clear and which data visualizations were likely to be more effective in the field.”

Another concern with having a core community of workers is that they eventually can become “expert” subjects. Whereas college undergraduates may spend a year or two participating in several studies across many psychology disciplines, MTurk subjects, on average, participate in dozens of studies, sometimes simultaneously or over very short periods of time.

Conducting dual online and offline studies also allows researchers to replicate their own work within different populations before seeking publication.

There also are choices for researchers who want to conduct Internet research while avoiding MTurk altogether, as several alternative services have sprung up to fill a perceived need. A site called Prolific Academic advertises high-quality, diverse, and naive subjects. The service requires that researchers pay subjects a minimum of \$7.50 per hour. Prolific Academic was founded by Ekaterina Damer (University of Sheffield, United Kingdom), a doctoral student in psychology, and Phelim Bradley (University of Oxford, United Kingdom), who is pursuing a PhD in genomic medicine and statistics.

“We designed Prolific specifically with academic research in mind,” Damer said. “Prolific has a pool of more than 24,000 participants around the world, so you can test for effects of different cultures.”

Qualtrics, the popular survey software, also advertises its Qualtrics Online Sample service. Enter the number of subjects needed, the length of the survey, and several other details, and a researcher is on his or her way to beginning data collection. The site SurveyMonkey offers a similar service called SurveyMonkey Audience. Other choices are detailed in a 2015 preprint by Eyal Peer (Bar-Ilan University, Israel) and colleagues.

Of course, the cost of these made-to-order samples varies. According to personal communication with the Qualtrics marketing team, 100 subjects participating in a 20-minute study would be “looking towards the low thousands.” In other words, none of these options is likely to beat MTurk in terms of affordability.

Proceed With Caution

Researchers’ mixed views about MTurk are captured in a 2015 special section in the journal *Industrial and Organizational Psychology*. Richard Landers (Old Dominion University) and Tara Behrend (The George Washington University) led the discussion with an article emphasizing that all convenience samples, like MTurk, have limitations, and that scientists shouldn’t be afraid to use these samples as long as they consider the implications with care. Among other recommendations, the authors cautioned against automatically discounting college students, online panels, or crowdsourced samples, and warned that “difficult to collect” data is not synonymous with “good data.”

While other researchers warned about repeated participation, motivation, and selection bias, APS Fellow Scott Highhouse and Don Zhang, both of Bowling Green State University, went as far as to call Mechanical Turk “the new fruit fly for applied psychological research.”

Finally, as Buhrmester noted, the convenience of online studies affects not only how subjects behave in studies, but also how experimenters conceive of the studies in the first place.

“There’s the issue of whether MTurk has become *too* popular among researchers,” he said. “I wonder how often researchers have chosen to design a study around what’s possible on MTurk rather than what would be most ecologically valid. I also wonder whether researchers are doing as much as they can to provide truly rewarding research experiences.”

Despite those concerns, Buhrmester believes MTurk has strong potential.

“At the end of the day,” he said, “it’s in everyone’s best interest for the MTurk community to grow and prosper.” ●

Tara Behrend will be speaking at a Cross-Cutting Theme Program on “Advancing Psychological Science Through Technology” at the 2016 APS Annual Convention, May 26–29 in Chicago, Illinois.

References and Further Reading

- 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, 3–5. doi:10.1177/1745691610393980
- Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26, 1131–1139. doi:10.1177/0956797615585115
- 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, 142–164. doi:10.1017/iop.2015.13
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184–188. doi:10.1177/0963721414531598
- Stewart, N., Ungemach, C., Harris, A. J. L. X., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment & Decision Making*, 10, 479–491.

Types of Sampling Procedures

1. **Uncontrolled**—researcher has no control in the selection of respondents
 - e.g., magazines, radio and TV call-ins
 - usually a very small sample—about 2%
 - usually biased in favor of more vocal individuals motivated to respond

2. **Haphazard sampling**—sampling procedure where the researcher may have some control over selection into study but it is still basically a hit-and-miss method for selecting participants
 - e.g., TV station sending crew out to interview people on the street for the evening news with instructions to include at least 2 women, 1 teenagers, and 1 person in a business suit

3. **Probabilistic sampling**—sampling procedures in which the researcher makes an effort to assure that each person in the population has an equal chance of being represented
 - a. **Simple random**—sample chosen from an entire population such that every member of the population has an equal and independent chance of being selected

 - b. **Stratified random**—sample is chosen to proportionally represent certain segments in the larger population

 - c. **Cluster**—sample is selected by using clusters or groupings from the population
 - e.g., sampling every student in 10th class rather than every 10th student (simple random)

Margin of error