would not, then students are appropriate.[21] Unless the phenomenon or process is different for students than for others, the findings should be valid.[22]

Compliance With Authority

Because following instructions and pleasing authority is key to doing well in college, students may be more likely to try to give researchers the outcome they are looking for.[23] This phenomenon was covered in chapter 2 under demand characteristics. Recall from that chapter how Stanley Milgram rejected the use of students for his studies on obedience to authority for this very reason. Students also may more easily detect a study's hypothesis because of their experience participating in research,[24] which leads to more biased responses.[25] Others have shown that subjects who are savvy about research can deliberately control their responses and undermine the results.[26] Researchers should always conceal the hypothesis and use measures that reduce hypothesis-guessing from all kinds of subjects.[27] More about these kinds of measures will be covered in chapter 9. Also, students' regular interactions with authority figures (professors) may make students respond differently than nonstudents. One researcher testing two variations of stress reduction treatment found that one treatment worked on a sample of students but had no effect on nonstudents, and the other treatment worked on people from the community but had no effect on students.[28] Greater compliance also can argue in favor of students as subjects under some circumstances, which may result in more accurate answers and less error from nonresponses.[29]

For Review-No Commercial Use(2023)

Weaker Sense of Self

Having a weaker sense of self may be true not just of college students but also of other eighteen- to twenty-one-year-olds outside of college, as young people are typically in the process of discovering who they are at this stage of the life cycle. For the purpose of research, however, this contributes to more volatile opinions and being more easily swayed by persuasive arguments.[30] This argues against the use of students in studies of attitude change, for example.[31] Again, researchers should consider whether this characteristic is related to the outcome variable and, if so, steer away from students as subjects.[32] Basil gives an example of studies of attitude change as being less appropriate for a sample of students; however, the process of opinion formation may not be different, and so students could be appropriate.[33]

Myriad Other Differences

College students are systematically different from others on many fronts; for example, they are likely to use social media more and traditional media less,[34] and are less conservative

and dogmatic.[35] As with all research, careful consideration of a sample's systematic differences must be taken into account. When a study makes no claim of inferring to a larger population, and the outcome variable is not influenced by their differences, then college students are usually appropriate.[36] Evidence for or against differences potentially biasing a study should be drawn from theory, existing evidence, and logic. Without such evidence, the researcher should proceed with students if there are good reasons to use them instead of others.[37] Studies with student samples have the ability to add to our knowledge, and that should privilege their use when no objections can be confirmed.

Homogeneity

One argument *for* students as subjects that is not based upon it being easier or cheaper for the researcher is that students are homogeneous—that is, they are similar to each other on several characteristics. Homogeneity of subjects in experiments aids internal validity by helping control for confounds—differences in subjects' age, education, use of technology, or whatever else cannot account for differences in outcomes if they do not vary.[38] Possible contaminants are minimized and the variability of measures is decreased, which helps reject the null hypothesis and identify when a theory is false.[39] In the pursuit of equivalence in treatment and control groups, as we saw in chapter 7, having subjects who are similar on variables related to the outcome is key. Students are more homogeneous than nonstudents on some variables[40] but vary as much as the general public on others.[41]

For Review-No Commercial Use(2023)

Students are similar to each other in age, education, media use,[42] life experiences, political engagement,[43] and political knowledge,[44] for example. When it comes to political knowledge and engagement, they tend to be lower than the average adult, so researchers should avoid conducting studies of political issues with students.[45] Likewise, students have little experience with paying taxes,[46] or interest rates and mortgages, and should be avoided as subjects in studies of these issues.[47] This argument applies to any set of homogeneous subjects; for example, people who are politically engaged, knowledgeable, and active do not resemble the average person.[48] Therefore, subjects drawn from political organizations, such as the League of Women Voters, provide the benefits of homogeneity but should not be generalized to others. In some fields such as cross-cultural research, for example, students' similar education levels give good estimates of representative samples.[49]

Basic Psychological Processes

The main argument for using students is when research is concerned with basic psychological processes that should be the same for everyone.[50] That is, when the process studied operates among all people, then results with students are generalizable.[51] If there is no

reason to believe the cognitive processes of students are different from nonstudents, then the use of student samples is appropriate.[52] Others go further to say that if the research aims to draw theoretical conclusions, such as for fundamental research and theory testing, then the sample does not matter.[53] Yet others have argued that even this is suspect because of demonstrated inabilities to generalize from students at one college to another,[54] or even from students in one major to another, and in different years of their schooling.[55]

Other research has identified the task subjects perform in the study as key to determining whether students are appropriate subjects. Analytical tasks with simple decisions are more appropriate for students than those with complex decisions and moral or ethical dilemmas.[56]

Guarding Against Bias With Students

There are some steps that experimentalists can take to reduce problems associated with the use of students as subjects. Studying students when they are the population of interest is, of course, appropriate[57]—for example, in studies of teaching techniques or when they are the target market in advertising or consumer research.[58]

For Review-No Commercial Use(2023)

Researchers should also compare studies of students to nonstudents to see if there are differences.[59] A good place to start is with the literature; much can be found that demonstrates students are no different from others in many fields, such as accounting, nursing, and education, among others.[60] But there is also much that says students are not a good proxy for other populations, and some of it contradicts findings in the same fields.[61] This is another instance where the subject of chapter 3, knowing the literature, is crucial. Researchers should undertake at least a cursory review of the empirical evidence expressly for the purpose of discovering what it has to say about student samples.

Student samples also can be compared statistically to population data using percentages and a technique known as *average variability* or *variance of the mean*, which adds knowledge of expected variability.[62] (Formulas can be found in Cochrane 1977.)[63]

Using a combination of students and subjects from another population and then doing a statistical analysis to demonstrate if there are differences, which are reported in the text or a footnote, is another good approach.[64] When no differences are found, that helps assure readers that the student subjects are not inappropriate for this particular issue. If there are differences, then the study has shown that effects may be contingent on sample characteristics.[65] Furthermore, using different sources for samples within the same study is key to both internal and external validity.[66]

Researchers can also replicate their own studies (as well as those of others) with a nonstudent sample to see if it makes a difference.[67] As Lindsay and Ehrenberg say, "If a

study is worth doing at all, it's worth doing twice."[68] Some have called for a minimum of two within-study replications with qualitatively different subjects as a requirement for publication in top-tier journals.[69] When conducting replications, do heed the advice in chapter 5 in order to extend the findings in some meaningful way.

Another way to ensure student subjects do not confound results can be to use statistical controls or covariates. These should be variables that are related to the outcome variable or are very different in students than in others, such as demographics of age and education.[70] The differences between students and others are not problematic as long as they can be predicted and controlled.[71] Report these in the paper and discuss findings honestly if they did confound the results.

Finally, some[72] recommend testing for prior knowledge of the theory and method used, as well as experience with research in general, in order to determine the likelihood of hypothesis guessing.

As pointed out in chapter 5, it is best to use subjects that are as similar as possible to those being studied; if the research is concerned with some phenomenon about teaching, and teachers can be reasonably expected to agree to be subjects, then that is better than using students. However, using education majors can be a reasonable proxy, as they will be teachers in the foreseeable future. As Basil said in the quote opening this chapter, a person is no less qualified to be in a study the minute before receiving a diploma than the minute after. For example, an argument can be made that undergraduates at elite journalism schools already had significant experience with student media before being admitted to serve as a proxy for journalists.[73] Others have used students in the master of public administration program who had previously been professionals in the field.[74] The real concern for experiments is whether the sample is appropriate to what is being studied, and less about how it was drawn.[75] Applying the yardstick for surveys and other methods is not an appropriate way to measure the quality of subjects in experimental designs.

For Review-No Commercial Use(2023)

Finally, the paper should always discuss why students, or any type of subject for that matter, are appropriate for the study.[76] Peterson says this should be justified by the theoretical relevance of the subjects to test the specific research questions.[77] In some fields, this is rare; for example, Marriott found that in ten tax discipline journals over twenty years, 80% did not provide justification for their sample or acknowledge any limitations associated with the use of students.[78] In marketing and consumer research, 63% ignored limitations of the sample, according to one study.[79]

It is also important to limit generalizations from studies that use student samples, or provide any policy changes or practice recommendations to professionals from such studies.

For example, extrapolating findings from students to the tax-paying public or making tax policy suggestions is not appropriate, although studies routinely do this.[80] Instead, conclusions should not focus on "people," "individuals," or the professionals the study is aimed at, but on "participants in this study," "these subjects," or "the professionals in this study."

This is also a good place in the paper to reiterate the purpose of experiments—to establish causality, not generalize to a population—as explained in chapter 5. Much space in this chapter is devoted to the use of students as subjects, for it is one of the most prominent issues in experimental studies. The next section discusses other types of samples for experiments and also the objections that have been raised to them.

Using a sample that is not comprised of students does not automatically solve all a researcher's problems. Many of the issues described earlier also apply to other types of samples. In general, convenience samples are biased because subjects volunteer. They do this for a reason, perhaps because they care deeply about the subject being studied or because they are motivated by the incentive, whether it is money or something else.[81] While there is no perfect sample, even one randomly drawn from a population, there are other good sources of subjects for experimental studies. Next is a discussion of some of the most popular sources.

For Review-No Commercial Use(2023)

AMAZON'S MECHANICAL TURK

One of the newest sources of subjects for experiments is Amazon's **Mechanical Turk**, an online crowdsourcing platform where people are paid to perform various tasks that computers cannot.[82] Amazon developed MTurk, as it is known, in 2005 for its own use[83] but opened it up to others who needed the services that only human beings could provide, such as writing book summaries, comments, or captions, identifying items in photographs, transcribing audio and even grocery receipts,[84] among other things.[85] Amazon calls these **Human Intelligence Tasks**, or **HIT**s.[86] One of the most unusual tasks is having workers look for missing persons in satellite images.[87] No one has yet been found this way. The name Mechanical Turk comes from a chess-playing machine popular at eighteenth-century carnivals designed to look like a Turkish sorcerer. In reality, a small man was hidden inside who played (and usually beat) the opponents—thus, the name of the platform designed to hide human workers.[88] In 2010, social scientists discovered it as a means for recruiting people to participate in surveys and experiments.[89] Advantages include that it is easy, affordable, and potentially representative of the population.

Is MTurk Representative?

The early criticism of MTurk subjects was that they were younger, better educated, and had more females than the population.[90] But that has been changing, with demographics growing more representative of the U.S. population.[91] Now, many studies show that MTurk samples are comparable to U.S. workers, standard Internet samples, and other subject pools,[92] and are more diverse than students,[93] especially when it comes to young Asians and young Hispanic females.[94] But not all studies show this; in one study, students were actually more representative of the population than MTurk workers.[95] Other studies compare representativeness of MTurk workers to international populations.[96]

As MTurk has grown in popularity, so have the challenges to its external validity.[97] Exploring differences between MTurk and other sample sources has become almost a cottage industry. For a review of some of this literature, see More About box 8.2, Research on MTurk's Representativeness.

As with using students as subjects, researchers should think critically about using MTurk workers for samples, considering whether they are different from other subject pools and the population of interest.[110] Also, relying on one type of sample, whether it is students, MTurk workers, or subjects from any other source, brings drawbacks of its own.[111] Using a variety of sources within the same study is one important way that greater external validity can be achieved in experiments.[112]

For Review-No Commercial Use(2023)

Advantages and Drawbacks

MTurk has gained widespread popularity in the social sciences because of its benefits, but it has some disadvantages as well. In addition to a large general population, low cost is probably one of the biggest reasons researchers use it. Workers are paid by the task. Some researchers have paid between sixty-one cents and seventy-five cents for eight to twelve minutes of work.[113] Workers with high approval ratings from other **requesters**, the term MTurk uses for those who post HITs, command higher pay. For example, workers with 95% approval ratings were paid $3 in one study.[114] The average hourly wage works out to between $1 and $1.38.[115] Amazon also tacks on a percentage for its own fee. Even at the higher rates, MTurk pay can be considerably less than what is required for students and adults. The savings achieved by using MTurk samples can be used to run more subjects or add more conditions.

Recruiting subjects and collecting data are also fast and easy in MTurk. Once the HIT is posted, hundreds of thousands of workers are automatically notified. Contrast this with going from class to class to pitch a study, or spending several days in a library or coffee shop soliciting subjects. Data collection is equally swift. Researchers report gathering enough subjects in the span of a few hours to less than a day,[116] with one

MORE ABOUT . . . BOX 8.2
Research on MTurk's Representativeness

Much research revolves around the question of how representative of the population MTurk workers are. This has important implications for the generalizability of studies that use them as subjects, especially for survey research but also to experimental work. Here is a brief overview of some of the research on this topic, both pro and con. As more work is always being done, readers should check for more and newer literature.

Bartneck and colleagues found that subjects recruited through MTurk were different from those recruited on campus or online, but declared that the difference was of "no practical consequence."[98]

Huff et al. compared MTurk with the Cooperative Congressional Election Survey and found strong similarities, with a maximum difference of less than seven percentage points on variables including age, gender, race, and political, occupational and geographic information. They say, "There are strong reasons for researchers to consider using MTurk to make inferences about a number of broader populations of interest," and encourage researchers not to automatically dismiss a study because it used MTurk subjects.[99]

However, Levay et al. replicated questions from the American National Election Studies (ANES) series in 2012 with MTurk workers and found significant differences. These were reduced by the use of covariates for nine variables—age, gender, race and ethnicity, income, education, marital status, religion, ideology, and partisanship. The authors conclude that "MTurk respondents do not appear to differ fundamentally from population-based respondents in unmeasurable ways"[100] and recommend including the nine variables as covariates. For experimentalists, they say that "representativeness of MTurk does not threaten experimental inferences in most cases"[101] but caution against unmeasurable variables that might also differ.

For Review-No Commercial Use(2023)

Similar conclusions were reached with an earlier ANES from 2008–2009 by Berinsky et al., who found that MTurk subjects were less representative than ANES subjects or those from Internet panels but more representative of the U.S. population than in-person convenience samples.[102]

When it comes to individual differences in demographics and political variables, MTurk subjects can be different from the U.S. population.[103] Yet researchers still claim that conclusions can be drawn from MTurk samples that are better in many cases than samples of students, as long as one understands the differences.[104]

Comparisons are not limited to demographic similarities though; research also examines the type of task in the experiment and outcome variables for differences between MTurk and other subjects. For example, Paolacci et al. and Birnbaum found only slight differences in MTurk subjects and others in decision-making experiments.[105] Buhrmester et al. found no meaningful differences between MTurk and other subjects on various psychometric tests.[106] Bartneck et al. found no differences in the outcome of studies that used subjects from MTurk, online services, or students. They say that MTurk is more "suitable for studies on the general population rather than specific subgroups."[107] Crump et al. found mixed results with a host of tasks including subliminal priming and learning tasks.[108] And Krupnikov found that MTurk subjects produced results significantly different from students and a nationally representative sample of adults when they were asked to read articles.[109]

While all these conflicting results may seem confusing, the key is to know what individual variables, outcomes, and tasks MTurk workers are different from the population on, and then act accordingly.

study receiving between 5.6 and 40.5 responses in an hour.[117] Furthermore, researchers do not need to be present or reserve a lab. And there is always a steady supply of MTurk workers, unlike the ebb and flow of students across the semester and breaks between them.[118]

These same benefits come with drawbacks, however. The low pay causes workers to try to complete a task as fast as possible, which can result in them not paying attention, pressing the same response button repeatedly, or just randomly checking off answers. Subjects taking experiments online, whether on MTurk or any other platform, have significantly more distractions than subjects in the lab.[119] To help control for that, an **attention check** question or two can be incorporated. For example, at a couple of points in the study, a factual question about what subjects just read can be asked, or an instruction can be inserted to provide a certain answer to the next question, no matter what their actual response would be. For example, "To the following question, please respond by selecting '4' regardless of what your answer would be." Make it clear in the instructions, consent form, and at the beginning of the study that the study includes questions designed to ensure they are paying attention and that failure to answer these correctly will result in not being paid. Workers are keenly interested in getting paid, not only for the money but because not getting paid results in bad ratings that can reduce the number of higher-paying HITs they are eligible to do in the future. Questions about details of the stimulus also can be asked at the end.[120] In order to help reduce response error, it is also recommended that researchers visually check for patterns such as repeatedly pressing the same button, or use pairwise agreement for the same question worded differently and other mechanisms to detect subjects rushing through a questionnaire.[121] Studies have shown that subjects in MTurk and traditional surveys are no different in terms of attentiveness.[122] Workers can be identified by their Amazon ID to ensure they do not complete the same study more than once.

For Review-No Commercial Use(2023)

Paying more seems to correlate with more completed tasks,[123] but quality of data is not related to rate of pay.[124] Turkers say they participate in research for entertainment as well as pay.[125]

Other disadvantages include a researcher's inability to control the environment of MTurk subjects the way they could in a laboratory.[126] In addition, gathering demographics and measuring response times require additional software such as the Qualtrics survey platform or AdobeFlash because the Amazon platform does not support this.[127] MTurk can be used to both recruit subjects and also host the questionnaire directly, or the questionnaire can be hosted using a survey package like Qualtrics, SurveyMonkey, and others.[128]

These drawbacks may easily be outweighed for researchers with few students, such as those at small colleges, without subject pools, or in nonacademic settings.[129]

OTHER SUBJECT SOURCES

Other more traditional sources for obtaining subjects can be more expensive but usually do not come with the concerns of the student and MTurk samples described earlier. These include **opt-in participant panels** from commercial survey and market research including Qualtrics, Survey Sampling International, and Knowledge Networks, now known as GfK KnowledgePanel. Others include SurveySavvy, Harris Poll Online, YouGov's PollingPoint Panel, Polimetrix, and Survey Spot. Some researchers have used Craigslist. There are many more, but these seem to appear regularly in published academic papers. These firms have a pool of people who have agreed to take studies for compensation and choose which research projects they will participate in. Some also offer the ability to target a sample that is representative of a population or has certain characteristics—for example, registered voters.[130] Costs depend on the number of questions, type of subjects, and other details. Researchers report paying as low as $3 to $5 per subject and as high as $15 per person.[131]

When a survey experiment is conducted using subjects from a defined population, such as that provided by opt-in data platforms, some disciplines and journals ask for a response rate to be reported. In fact, the American Political Science Association's (APSA) experimental research section's standards committee guidelines specify that response rate should be provided for every experiment that uses survey data collection methods,[132] and even for college student samples.[133] There is debate about whether this is appropriate or necessary, with some researchers saying that it is meaningless if the study does not make any claims to generalizability.[134] Currently, not many social sciences disciplines, with the exception of political science, require response rates for survey experiments; however, researchers should follow the standards and guidelines in their field and journal they are submitting to.

For Review-No Commercial Use(2023)

A more traditional approach is to recruit subjects through advertisements in newspapers,[i] flyers posted around town, and mailing lists. These methods have been applied to new media, with social media and online forums now sources for subjects. For example, one study used a LEGO forum on the Internet for a study about judging emotional expressions that manipulated the faces on LEGO toys.[135] The authors noted the potential bias in that the subjects were homogeneous in their interest in and experience with LEGOs. Another group of researchers developed a method for recruiting from social media that allows for targeting subjects around specific themes, such as the politically engaged—a particular weakness of student samples.[136] It appeals to online opinion leaders, such as

[i]Qualtrics provides respondents for research and also has survey software that can be used to administer a study regardless of whether subjects are drawn from the firm's pool. Researchers can use the software without using subjects from the pool.

bloggers and discussion forum moderators, to recruit subjects. This type of recruiting is especially good for studying a targeted sample. For political scientists, for example, being able to find politically sophisticated subjects, who make up a small percentage of the overall population, can give more accurate estimates of effects while improving external validity. Another example is a study that uses subjects from chat rooms for hate groups to study the types of messages that trigger aggression.[137] The authors write, "This insight would not be possible through the use of nationally representative probability samples, conventional student samples, or other online opt-in methods such as MTurk.[138] (See Study Spotlight 8.3 for more about a study of social media sites as sources of subjects.) Researchers can also enlist the help of organizations for specific subgroups—for example, public relations professionals who join the Public Relations Society of America.

Other researchers have found subjects in the staff of their schools,[139] lists of jury pools, which are drawn by randomly sampling citizens and are more heterogeneous, and election rolls.[140] (See Study Spotlight 8.3 for more about a study of jury pools as sources of subjects.) Some researchers are able to include experimental questions on a large-scale survey for free—for example, the Time-Sharing Experiments for the Social Sciences—while others, such as the Cooperative Congressional Election Study, can be costly.[141]

For Review-No Commercial Use(2023)

Another advantage with this recruitment strategy is that subjects are homogeneous, reflecting shared interests, issue attitudes, beliefs, and values. They are also the types of people who are likely to self-select to be exposed to the types of messages being studied, overcoming the problem of using natural settings that artificially expose people to messages that they would not otherwise see.

RECRUITING

Subjects also can be recruited in public places such as parks and outdoor cafes. One researcher sat at tables outside coffee shops and other businesses, with the owners' permission, and recruited passersby with gift cards to the establishment.[148] Another found that fire departments were receptive to researchers' visits, as the firefighters had a lot of free time when waiting for calls. Subjects can also be recruited through social organizations such as Lions Clubs and church groups who invite speakers to their meetings. After having members take the study with paper and pencil for fifteen to twenty minutes, researchers can provide a fascinating talk about their research. This approach also has been used in newsrooms with ethics research; managing editors are usually receptive to having a professor give their journalists an ethics miniworkshop. Other strategies can be to set up tables at conferences to recruit working professionals—for example, at a conference for Internet domain name investors.[149] No doubt the creative display—an open briefcase full of money—helped attract subjects. When subjects need to represent a particular population—for example, working professionals—this

STUDY SPOTLIGHT 8.3

Jury Pools and Social Media as Subject Sources

Murray, G. R., C. R. Rugeley, D-G. Mitchell, and J. J. Mondak. 2013. "Convenient Yet Not a Convenience Sample: Jury Pools as Experimental Subject Pools." *Social Science Research* 42 (2013): 246–253.

Instead of relying on college students, these authors found another convenient, diverse, and inexpensive source of experimental subjects—people who have been summoned for jury duty. This paper discusses practical concerns such as gaining access and administering the experiment, and evidence including response rates, sample quality, and representativeness.

Other researchers have used prospective jurors, including Niven,[142] Sigelman et al.,[143] Terkildsen,[144] and Lewis et al.[145]

Some concerns with the use of jurors as study subjects include gaining the cooperation of court officials and judges. These authors, who have been conducting experiments with jury pools since 2009, report no problems and say that the good relationship the county has with the university helps. Not all have found this to be the case; for example, a nationwide study that used jury pools had only a 15% cooperation rate.[146]

Other issues include concerns by court administrators about increased workload and disruption of court proceedings by judges. Judges also are concerned that jury members are treated fairly and the study does not bias them. Other concerns aligned with Institutional Review Board (IRB) requirements, such as that subjects be free to not answer questions or stop the study at any time.

For Review-No Commercial Use(2023)

Judges are concerned about the content of the study, wishing to avoid anything to do with legal issues; political topics were OK. Judges review the study and approve it, and these authors have only had one question removed in all their experience.

Procedurally, the researchers use paper-and-pencil formats but say that handheld electronic devices might also be possible. They suggest limiting studies to ten to fifteen minutes because of potential jurors' limited free time. They administer the study when jurors first arrive and are waiting for the selection process to begin. They place clipboards with pencils attached, the cover letter, and instructions and questionnaire on seats before jurors arrive. They spent $350 on clipboards and incentives. One researcher should be present to monitor the study and answer questions. They have not found it necessary to offer incentives and estimate the cost at fifty-seven cents per subject—comparable to the cost on MTurk—which includes photocopying and supplies but not transportation to the courthouse.

In these researchers' county, jury pools are called twice a week with approximately seventy-five to 200 people in the pool. They estimate an average of 400 completed studies per month in a county of 100,000. Their response rate was 79% among potential jurors who showed up; adjusting for all those summoned, the response rate was between 39% and 68%.

Because jury pools are drawn from voter registration and driver's license lists, among other sources, jury pools are diverse and representative of the community. These authors compared jurors to a sample of students and a telephone survey using the same study. You can read their results in a table. They conclude, "Overall, we contend that jury pool data equal or exceed the alternatives under consideration on cost effectiveness, response rates, diversity of respondents, and representativeness."[147]

(Continued)

(Continued)

Cassese, E. C., L. Huddy, T. K. Hartman, L. Mason, and C. R. Weber. 2013. "Socially Mediated Internet Surveys: Recruiting Participants for Online Experiments." ***PS: Political Science & Politics*** **49 (4): 1–10.**

This study reports on a method the authors have used in six studies to recruit experimental subjects using social media, which they call the Socially Mediated Internet Survey (SMIS) method. It works by asking bloggers, discussion forum moderators, and Facebook and Google+ page administrators—all of whom are "central nodes" in the social network—to help recruit readers and visitors to their sites as study subjects.

Among the advantages of this method is that special subpopulations, such as political sophisticates and activists for a specific social issue, can be recruited. Political scientists, the discipline of these researchers, are often interested in highly engaged, knowledgeable, and politically active citizens, but they are not widely found in the general population. Using traditional recruitment methods would result in many subjects that do not fit the requirements of political sophisticates. With this method, they can be specifically targeted. For example, politically involved people interested in immigration issues can be recruited via the Stand With Arizona (and Against Illegal Immigration) Facebook page, which had 600,000 followers in early 2013. Researchers of various disciplines could use this; for example, a study of health messages about vaccines could recruit subjects via Facebook pages Vactruth.com and Anti-Anti-Vaccine-Campaign.

Researchers use the SMIS method by first identifying social networks that appeal to the subpopulation of interest, and then by appealing to bloggers and page administrators to endorse their study and solicit participation from their readers, contacts, and followers. They suggest Technorati (www.technorati.com) and lists of blogrolls (e.g., BlogHer.com for a directory of women bloggers) to identify blogs and forums. Because the request to participate in a study comes from a known opinion leader rather than an unknown researcher, the personal connection helps increase the likelihood of participation.

For Review-No Commercial Use(2023)

The authors give advice for gaining trust and cooperation by contacting the social mediator to obtain permission and answer questions about the project, noting that IRB approval is required for this contact. They give a link to an online supplemental appendix that contains the materials they used. In the six studies reported in this paper, about one of every eight bloggers the researchers contacted agreed to promote the study. They say they secured an average of 1,569 subjects per study, considerably more than they would have been able to recruit via student subject pools. Each social mediator recruited an average of 104 subjects. And unlike MTurk, these subjects were volunteers, and no incentive or payment was required.

The researchers compare their SMIS samples to samples from MTurk, students, and 2008 ANES subjects, and conclude that, while they are not representative, they are more diverse than students demographically and geographically.

approach works well. However, researchers should realize that in the cases of social clubs and organizations, the sample will be limited to self-selected groups who have time or are interested in the topic and share particular characteristics that may not be representative of the population—for example, retirees or women not employed outside the home. For this reason, sampling from a variety of sources to avoid systematic bias can be a good strategy.

INCENTIVES

While some researchers have found they do not need to offer incentives,[150] many find that even a small token is important to convey to subjects that their time and effort in the research

is valued.[151] Books, pencils, folders,[152] and postage stamps[153] are also used in addition to financial incentives. Gift cards can be more cost effective than cash; for example, more people may decline to participate for $10 in cash than for a $5 gift card. It may work best to describe the incentive as "a gift card to (name of coffee chain)" and leave out the amount; rarely does anyone ask how much it is for. It is a good idea to call ahead to make sure the store has enough cards on hand or if purchasing cards in bulk must be done at corporate headquarters.

Another approach is to tell participants they will be entered into a lottery for a drawing of a larger prize, such as an iPad. Two researchers raffled off a book and got twenty-nine students to participate.[154] Researchers also can raffle off a limited number of $10 gift cards, meaning they can buy fewer cards than the number of subjects.[ii] Not all IRBs will allow lottery incentives, so check first.

In addition to explaining the rationale for the use of the subjects in the study, researchers should also report the incentives, if any, and other details that will help readers evaluate the study. In the field of political science, the APSA's experimental research section's standards committee has codified a list of guidelines that includes who was eligible to participate, the dates of recruitment, when the study was conducted, and the number of subjects assigned to each group, among other things. For a complete list of guidelines, see Gerber et al. 2014 and 2016. Other disciplines and also individual journals may publish their own reporting standards. The most widely used standards are found in the *Publication Manual of the American Psychological Association.*

For Review-No Commercial Use(2023)

No doubt there are many other good sources of subjects for experiments and ways to recruit and incentivize them, limited only by one's imagination. When reading studies, pay attention to who the subjects are, where they came from, and how they were recruited. Also note the number of subjects used, which is the topic of the next section.

SAMPLE SIZE AND POWER

"How many subjects do I need for my experiment?" is one of the first questions students ask. The good news is that experiments usually can require far fewer subjects than methods such as surveys. But the answer to the question is not quick or simple.

The topics of sample size and power are intertwined. **Sample size**, or the number of subjects in a study, is referred to as *N* or *n* in experiments, with the capital *N* representing all subjects in the study and lower-case *n* representing the subjects in one of the conditions. Estimating the sample size while in the process of designing an experiment is important

[ii]Many researchers self-fund their studies. Check with your department, college, and university for sources of funding, as well as outside grantors.

to ensure that statistical conclusions are accurate and reliable. If the sample is too small, the statistical test may not be significant because of low power, not because the hypothesis is incorrect.[155] If the sample is too big, the researcher will have spent time and money for little extra benefit. Thus, it is important to have the number of subjects be just right.[iii]

This is where statistical **power** comes in. Power is the probability of supporting a hypothesis, or, technically, rejecting a false null hypothesis when the alternative hypothesis is true. That is, power is the probability of achieving statistical significance when an effect actually exists. Power also is the concept behind **Type 2 errors**, which occur when the null hypothesis is not rejected but should be; that is, the alternative hypothesis should be supported, but it is not. This is indicated when the *p* value is not lower than .05. When a significance level of .05 or less is reached, there is more reason to be confident that the results are "real" and not due to chance. When statistical significance is not achieved due to a Type 2 error, besides being extremely disappointing for researchers, this means that some treatment that may actually work is now reported as not having an effect. Important interventions may never be adopted in practice.

Researchers tend to focus on **Type 1 errors**, which are the opposite of Type 2—that is, when the null hypothesis is rejected, but it should not be.[156] But Type 2 errors are also common in published research.[157] Researchers need to focus on both reducing errors and increasing power, and on guarding against Type 1 and Type 2 errors.[158]

For Review No Commercial Use (2023)

One of the easiest ways to increase power, thereby reducing Type 2 errors, is to increase the sample size.[159] Larger samples mean more power because they more accurately represent the population from which they were drawn.[160] For an example of how this works, assume the entire population consists of one million people. It is difficult to measure all one million people, so scientists draw a sample from the population instead and estimate the mean on some variable. If the sample consisted of ten people, it would be possible to draw people with especially low or high scores. Increasing the sample to 100 increases the probability of drawing people with more moderate scores. The sample mean based on 100 people is closer to the population mean than the one based on ten people.

Another important statistic in this calculation is the **standard error**, which estimates the average difference between the sample and the population statistic.[161] It is the standard

[iii] American Psychological Association, *Publication Manual of the American Psychological Association,* 6th ed. (Washington, DC: Author, 2009). The APSA guidelines also include reporting an intent-to-treat analysis, which is found more in the medical field than social science. It is beyond the scope of this book to explain this statistical analysis, but instead, readers are referred to statistics books and courses, and articles such as Sally Hollis and Fiona Campbell, "What Is Meant by Intention to Treat Analysis? Survey of Published Randomised Controlled Trials," *BMJ* 319 no. 7211 (September 1999): 670–674.

deviation of the sampling statistic. Standard errors tend to be smaller when samples are larger; thus, larger samples increase power. But because time and money are limited in most cases, it is usually not practical to increase the sample size infinitely. Consequently, researchers need to balance the need for larger samples with the resources available.

It is also important not to have *too many* subjects, which could result in detecting relationships among all variables.[162] If power is too strong due to an excessively large sample, then finding statistical significance is almost certain. While it may seem like a great idea to assure one's hypotheses are supported, the relationships may be so small that they are unimportant practically.[163] The findings would then be substantively unimportant, which is the topic of the next section on effect sizes. Readers, reviewers, and editors are becoming savvy at detecting when a study has supported its hypotheses mainly by packing it with too many subjects. In addition, it is more ethical to minimize potential harm by using the least number of subjects necessary.[164]

It is also important to know any specific issues in your discipline that might affect the size of a sample; for example, in advertising, a phenomenon known as **ad blindness** occurs where subjects subconsciously tune out ads.[165] Advertising researchers may need to sample as many as 30% more subjects just to account for purging those who are unable to provide meaningful responses (about 30% cannot correctly recall seeing it because they did not recall seeing it. In moral development research, about 8% of subjects fail questions designed to detect those trying to fake a higher score and are purged.[166]

For Review No Commercial Use (2023)

Effect Sizes

The term *power* also refers to a study's **effect size**,[167] which is a statistic that describes the strength of the relationship between the independent variable (IV) and the dependent variable (DV), or the amount of variance in the DV that is explained or accounted for by the IV. The effect size statistic serves as a complement to the statistical significance represented by the *p* value. Effect size statistics refer to the *practical* significance of the relationship. In other words, the *p* value tells whether the treatment worked; the effect size tells how well it worked. More About box 8.4 explains the different effect size statistics.

Effect sizes may or may not be related to the *p* value. For example, a highly significant *p* value may be accompanied by a low effect size statistic, indicating that there is a low probability of the treatment affecting the outcome by chance, but the size of the effect is practically meaningless. The best possible outcome is when the *p* value is significant and the effect size is large; this means that the treatment effect is not only real but also important. The effect of the treatment is meaningful in a practical sense.

MORE ABOUT . . . BOX 8.4
Effect Size Statistics

Different statistics represent effect sizes for different statistical tests. The effect size estimates the amount of variance explained or accounted for by the independent variables in an experiment. The first column contains the statistical test and the second the effect size statistic with its symbol and an explanation. These vary from 0 to 1.

ANOVA, MANOVA	Eta-squared, $\eta 2$
Tests differences between two or more groups. Multiple analysis of variance (MANOVA) tests multiple dependent variables.	Eta-squared measures the variance explained in the sample, not the population, so it overestimates the effect size. The bias gets smaller as the sample size increases.
	SPSS does not calculate this; it must be calculated by hand. http://www.theanalysisfactor.com/calculate-effect-size/
	Partial Eta-squared, η^2_p SPSS calculates partial eta-squared for analysis of variances (ANOVAs), but calls it eta-squared. It is the same value as eta-squared with one IV designs, but different for repeated measures designs. With more than one IV, partial eta-squared is the variance explained for one IV controlling for the other IVs. In other words, it excludes or "partials out" the variance explained by the other IVs. It is preferred when more than one IV is tested, as it is more conservative.
	Omega-squared, ω^2 Estimates the variance explained in the population. Limited to between-subjects designs with equal cell sizes. Less biased than eta-squared or partial eta-squared.
Correlation and Regression Tests the association or relationship between two or more variables; tests variables that predict an outcome variable.	***r*-squared, r^2 or R^2** The correlation between two variables is *r,* then the squared value of *r* is r^2. This represents the amount of variance in each that is accounted for or explained by the other. For example, if $r = 0.21$, $r^2 = 0.0441$. 4.4% of the variance of either variable is shared with the other variable. In multiple regression, it is expressed as R^2 and explains the variance accounted for by all the independent variables together.

For Review. No Commercial Use (2023)

ANOVA, MANOVA	Eta-squared, η2
Odds Ratio	**OR** This effect size measure is commonly used in medicine for binary variables or a dichotomous outcome, such as survival or not. It is expressed as, for example, the odds of a treatment success are three times higher than in the control group.
Chi-Square Test of difference between categorical variables.	**Phi coefficient, Cramer's V or Phi, φ** **Cohen's *d*** Examines the strength of the relationship between two variables; V may be used with more than two levels of a variable.

For Review-No Commercial Use(2023)

Effect sizes are important to report along with significance tests because this allows readers to see the effect of a treatment not confounded by sample size. Whereas statistical significance is heavily influenced by sample size, effect sizes are less so. For example, a sample of 500 may show high statistical significance but have small effect sizes. The reverse is also true; a sample of fifty may not show statistical significance, but the effect size may be large. Having both statistics reported in a study allows readers to decide for themselves the substantive importance of the treatment. Reporting of effect sizes is standard practice in psychology[168] and is considered good practice in other disciplines.[169] Many journals are now requiring it. Whereas in the past a study that failed to reach statistical significance was likely to be rejected,[170] nowadays, one that reaches significance is not guaranteed publication if it is accompanied by a very low effect size.

Reporting effect sizes merely requires the addition of the proper statistic in the results section at the end of the list of values showing the results of the significance tests. In the following examples, the effect sizes are underlined:

- "Using condition (action exemplar, inaction exemplar, and control) as the independent variable and negative affect as dependent variable, a one-way analysis of variance (ANOVA) reported a significant effect, $F(2, 192) = 9.24$, $p < .001$, $\underline{\eta 2 = .09}$."[171]
- "The impact on candidate 'image' evaluation under three issue conditions resulted in a significant finding ($F = 6.92$, $df = 6, 450$, $p < .01$, $\underline{\eta 2 = .16}$); the finding also was significant for candidate 'capability' ($F = 11.80$, $df = 6, 450$, $p < .01$, $\underline{\eta 2 = .25}$)."[172]

These examples used eta-squared because the test performed was ANOVA. Next is an example that uses r^2 for a correlation test:

> "H3 was supported: there was a significant positive correlation between the participants' liking of stories on the experimental website and civic engagement (r =.811, p < .001, r^2 = .66)."[173]

Cohen[174] categorized the relative effect sizes according to this convention, also called "T-shirt effect sizes,"[175] and used in most social science disciplines:

Cohen's Conventions for Effect Sizes	
Small	< 0.20
Medium	0.20–0.50
Large	> 0.50

While Cohen provides a general rule of thumb, the magnitude of an effect size depends on the context. 5 may be practically significant for some disciplines and 50 considered large.[176] As you read studies in your field, notice what the reported effect sizes are in order to get a sense of your discipline and the specific variables' effect sizes. In the planning stages of any experiment, it is important to determine a minimum effect size that would indicate a treatment or manipulation is useful or important.[177]

For Review-No Commercial Use(2023)

Sample size, effect size, power, and significance level are all connected; if you know any three of these, the fourth can be determined. How to do that with a power analysis is the topic of the next section.

Power Analyses

Determining sample size is the most common use of power.[178] Power estimations and analyses should always be done before data are collected, known as **a priori** power analysis. There are also rule-of-thumb methods for determining how many subjects a study should have, but free online power analysis calculators have all but replaced these. (See table 8.1 for Rules of Thumb for Sample Sizes.) All analyses of power are dependent upon the statistical test to be performed. For example, experiments that use regression analysis or structural equation modeling require more subjects than those that use t tests or ANOVA. So far in this text, we have concentrated on studies that test differences between groups, which are appropriately tested with statistics such as the t test when two

groups are compared, ANOVA for three or more groups, and MANOVA, which is a form of ANOVA used when there are multiple DVs. A separate statistics course is required to conduct these tests knowledgeably, and this book is not intended to replace that. But readers should be aware that the statistical test performed affects the power and thus the size of the sample needed.

TABLE 8.1 ⬢ RULES OF THUMB FOR SAMPLE SIZES WITH VARIOUS STATISTICAL TESTS

t tests, ANOVA, MANOVA	n = 30 per cell (Cohen, 1988)
Correlation, regression	n > 50 + 8 per IV to N > 104 + 8 per IV (Green 1991)[179] The number of predictor variables + 50 (Harris 1985)[180] 30 subjects per variable (Cohen and Cohen 1975)[181]
Chi Square	No expected frequencies below 5 (Howell 1997)[182]
Factor Analysis	300 subjects (Tabachnick and Fidell 1996)[183], 50 per factor (Pedhazur and Schmelkin 1991[184], Comrey and Lee 1992[185])

For Review-No Commercial Use(2023)

Notes: These sample sizes are for medium to large effects sizes.

Cohen[186] suggests a minimum power of 80%.

Comparisons of fewer groups/cells requires more participants to maintain power.[187]

Once upon a time, a power analysis had to be conducted by hand. More commonly today, researchers calculate power analyses using statistical software such as Stata and R, or online tools, one of the most popular of which is G*Power. Created by a group of German researchers, the software is free, available at http://www.gpower.hhu.de/, and is accompanied by tutorials.[188]

Before conducting a power analysis, researchers should examine previous studies that are similar to the one they are conducting and collect basic data including means, standard deviations, and effect sizes from them. Prior studies that researchers have conducted themselves also are useful for this purpose. For example, for an experiment that looked at whether journalists used different levels of moral judgment when story subjects were children than adults,[189] data from previous work on photographs and moral judgment were used to figure power and determine the sample size.[190]

In order to determine the sample size, *N*, at least three of the following four pieces of data are required:

- Alpha level, α—The convention in the social sciences is to use .05 or lower. A test will have more power with an alpha level of .05 than .01.
- Power—At least 80%, or .8 or higher, is recommended.[191]
- Effect size, *d*—Cohen's conventions, listed earlier, offer a rule of thumb. Read existing studies to determine what is typical for a study of a specific topic.
- Means and standard deviations—Obviously, means and standard deviations are not available on a study that has not yet been conducted, so these can come from previous research similar to the study to be conducted,[192] prior studies the researcher has done on the topic, or pilot studies. Meta-analyses are another good place for finding means from many different studies.

In the example of a moral development study, two groups were used—a treatment (stories about children) and control (stories about adults)—so *t* tests were specified in G*Power. A directional hypothesis was not predicted, so two-tailed tests were used. (If a direction had been predicted, for example, that animals would use higher levels of moral judgment for stories with children than adults—a one-tailed test would have been used.) The a priori mode was specified because the power analysis was conducted before the study was performed to discover how many subjects to use. The means and standard deviations from the previous study of photographs were inserted; these were 5.27 (1.18) for the treatment group (photographs) and 4.25 (1.50) for the control group (no photographs). For power, .8 was specified; researchers can vary this between .8 and .95 to see how it affects the sample size. The Alpha error probability should be .05, which is standard in the social sciences. When these data are entered, the G*Power command will calculate the results. In this case, G*Power indicated a total of 94 subjects, with 37 in each of the two groups. Because the means and standard deviations were used, it calculated an effect size of .75, which is considered large by Cohen's conventions. If you do not know or cannot estimate the means and standard deviations, then input the desired effect size instead. Vary it to see how it changes the sample size requirements. Increasing the effect size can increase the sample size dramatically; specify smaller effect sizes if those are common for the research topic.

For Review-No Commercial Use(2023)

To review, it takes three of the pieces of information to calculate the fourth—alpha level (.05) and power (.8 minimum) are already determined; examine the literature to determine either effect size or the means/standard deviations. Experiment with the G*Power analysis to see how many subjects it takes to get a larger effect size. Performing a power analysis can help you decide if you need to switch to a cheaper form of subjects, such as MTurk, or if you can afford a sample from the general adult population.

This illustration of a G*Power analysis is fairly basic; there is actually much more to it. For example, G*Power allows the use of F tests using one-way ANOVA, main effects and a one-way interaction, and ANOVA between-subjects or within-subjects, among others. This requires a more advanced understanding of statistics, and more in-depth G*Power tutorials are available from many sources. Some are listed some in the "Suggested Readings" section.

Reporting a power analysis can range from a sentence to all the details in an appendix. Some examples include:

- "An a priori power analysis was completed based on the effect size of $d = 0.65$ for self-referential encoding compared to semantic encoding as reported in Symons and Johnson's large-scale meta-analysis (1997). The power analysis indicated that with power at 95% and an alpha of .05, a sample size of 33 would be sufficient to detect an effect of self-referential encoding."[193]
- "Data were analyzed from a final sample of 20 participants (mean age = 19.70, 15 females) who completed the study procedures. This sample size was chosen based on a priori power analyses (see supplemental materials)."[194]
- "An a priori power analysis, conducted in G*Power 3, indicated that a total sample size of approximately 60 participants would yield acceptable power (i.e., $1\text{-}\beta > 0.80$) for testing moderate effect sizes."[195]

For Review-No Commercial Use(2023)

Observed Power

Finally, there is something called **observed power**, **post hoc power**, or **a posteriori power**, among other terms, that a few journals request researchers include when a statistical test was not significant.[196] This refers to a power test done after the study has been conducted, with the aim of understanding if the reason the null was not rejected could be attributed to the study being underpowered.[197] This is a controversial and complicated topic, the details of which will not be discussed here. To simplify, statistical tests that are not significant *always* have low power.[198] There is, in fact, a one-to-one correspondence between power and the p value.[199] Many studies inappropriately report low power as a possible explanation for failure to support a hypothesis.[200] This is especially true when reporting the observed power values provided by SPSS or G*Power.[201] Instead, post hoc power should be based on population effect sizes of independent interest,[202] or results should be described using p values, effect sizes, and confidence intervals.[203] As this is beyond the scope of this book, readers are instead referred to statistics books and courses.

The next chapter is concerned with the stimuli that subjects will receive in an experimental study and the instrumentation or observations that will be used to measure their responses.

Common Mistakes

- Using student samples without considering how they might be different from the population of interest
- Not providing a rationale for the use of subjects of any type
- Generalizing beyond the subjects in the study
- Not conducting a power analysis prior to executing the study to determine the number of subjects needed
- Not reporting effect sizes, power, alpha, and number of subjects for each group

Test Your Knowledge

1. Most experiments use which kind of samples?
   a. Random samples
   b. Representative samples
   c. Probability samples
   d. Convenience samples

For Review-No Commercial Use(2023)

2. Which of the following is NOT of concern when students are used as experimental subjects?
   a. Higher than average cognitive skills
   b. Homogeneity
   c. Compliance with authority
   d. Weaker sense of self

3. Which of the following is a disadvantage of using Amazon's Mechanical Turk for subjects?
   a. It is expensive
   b. Data collection is slow
   c. Subjects not paying attention to questions
   d. Subjects are not diverse

4. A Type 2 error is _____.
   a. The null hypothesis is *not* rejected when it should be
   b. The null hypothesis is rejected when it should be
   c. The null hypothesis is rejected when it should *not* be
   d. None of these

5. Power is related to _____.
   a. Sample size
   b. Effect size
   c. Significance level
   d. All of these

6. Which of the following is a problem related to a large sample?
   a. It raises the cost and time required to conduct the study
   b. Detecting relationships among all variables
   c. Finding significant relationships that are practically unimportant
   d. All of these

7. Effect size is _____.
   a. The amount of variance in the DV that is explained or accounted for by the IV
   b. The same as the significance level
   c. Heavily influenced by sample size
   d. Something that should not be reported in the study

For Review-No Commercial Use(2023)

8. In the social sciences, an effect size of .35 is considered to be:
   a. Very Small
   b. Small
   c. Medium
   d. Large
   e. Very Large

9. Estimating sample size with a power analysis before a study is conducted is called:
   a. A posteriori power
   b. A priori power
   c. Post hoc power
   d. Observed power
   e. Abuse of power

10. Power is not dependent upon the statistical test to be performed
    a. True
    b. False

Answers

1. d
2. b
3. c
4. a
5. d
6. d
7. a
8. c
9. b
10. b

Application Exercises

1. Do a power analysis using data from studies similar to the one you are working on. Use G*Power or some similar power calculator. A free download of G*Power, the manual, and a tutorial can be found at http://www.gpower.hhu.de/.
2. Write up the sampling portion for the methods section of the experiment you are creating. Follow the examples in this chapter and other studies that are similar to yours. Address the following: Who are you going to use and how do they relate to the target population, how many do you need, how will you recruit them, what incentives will you use, and any other topics pertinent to your study. Incorporate your power analysis. Justify the use of your sample. One to two pages.

For Review-No Commercial Use(2023)

Suggested Readings

On student subjects:

Basil, Michael D. 1996. "The Use of Student Samples in Communication Research." *Journal of Broadcasting and Electronic Media* 40 (3) (Summer): 431–440.

On effect size:

Levine, T. R., and C. R. Hullett. 2002. "Eta Squared, Partial Eta Squared and the Misreporting of Effect Size in Communication Research." *Human Communication Research* 28: 612–625.

On G*Power:

Mayr, S., E. Erdfelder, A. Buchner, and F. Faul. 2007. "A Short Tutorial of Gpower." *Tutorials in Quantitative Methods for Psychology* 3 (2): 51–59.

On Post Hoc Power:

O'Keefe, Daniel J. 2007. "Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses." *Communication Methods and Measures* 1 (4): 291–299.

Notes

1. Michael D. Basil, "The Use of Student Samples in Communication Research," *Journal of Broadcasting and Electronic Media* 40, no. 3 (1996): 439.
2. William D. Wells, "Discovery-Oriented Consumer Research," *Journal of Consumer Research* 19 (1993): 492.
3. Elizabeth Tipton, "How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations," *Journal of Educational and Behavioral Statistics* 39, no. 6 (2014): 478–501.
4. Basil, "The Use of Student Samples"; John A. Courtright, "Rationally Thinking About Nonprobability," *Journal of Broadcasting & Electronic Media* 40, no. 3 (1996): 414–421; Annie Lang, "The Logic of Using Inferential Statistics with Experimental Data From Nonprobability Samples: Inspired by Cooper, Dupagne, Potter, and Sparks," *Journal of Broadcasting & Electronic Media* 40, no. 3 (Summer 1996): 422–430; William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Belmont, CA: Wadsworth Cengage Learning, 2002).
5. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs*, 373.
6. Basil, "The Use of Student Samples."
7. Robert A. Peterson, "On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-Analysis," *Journal of Consumer Research* 28 (2001): 450–461; I. Simonson et al., "Consumer Research: In Search of Identity," in *Annual Review of Psychology*, ed. S. T. Fiske, D. L. Schacter, and C. Zahn-Waxler (Palo Alto, CA: Annual Reviews, 2001): 249–275.
8. Lisa Marriott, "Using Student Subjects in Experimental Research: A Challenge to the Practice of Using Students as a Proxy for Taxpayers," *International Journal of Social Research Methodology* 17, no. 5 (2014): 503–525.
9. W. J. Potter, R. Cooper, and M. Dupagne, "The Three Paradigms of Mass Media Research in Mainstream Communication Journals," *Communication Theory* 3 (1993): 317–335.
10. Joanne M. Miller, "Examining the Mediators of Agenda Setting: A New Experimental Paradigm Reveals the Role of Emotions," *Political Psychology* 28, no. 6 (2007): 696.
11. Elizabeth A. Bennion and David W. Nickerson, "The Cost of Convenience: An Experiment Showing E-Mail Outreach Decreases Voter Registration," *Political Research Quarterly* 64, no. 4 (2011): 861.
12. Renita Coleman, "The Effect of Visuals on Ethical Reasoning: What's a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass Communication Quarterly* 83 no. 4 (Winter 2006): 840.
13. Emily Thorson, "Belief Echoes: The Persistent Effects of Corrected Misinformation," *Political Communication* 33, no. 3 (2016): 466–467.
14. M. Kim and G. G. Van Ryzin. "Impact of Government Funding on Donations to Arts Organizations: A Survey Experiment." *Nonprofit and Voluntary Sector Quarterly* 43, no. 5 (2014): 916.
15. Hoon Lee and Nojin Kwak, "The Affect Effect of Political Satire: Sarcastic Humor, Negative Emotions, and Political Participation," *Mass Communication and Society* 17, no. 3 (2014): 315.
16. Renita Coleman, "Journalists' Moral Judgment About Children: Do as I Say, Not as I Do?" *Journalism Practice* 5, no. 3 (Summer 2011): 257–271.
17. D. O. Sears, "College Sophomores in the Laboratory: Influences of a Narrow Database on Social Psychology's View of Human Nature," *Journal of Personality and Social Psychology* 51 (1986): 515–530.
18. Ibid.
19. Basil, "The Use of Student Samples," 434.
20. Ibid.
21. Ibid.
22. Ibid.
23. Ibid.; Sears, "College Sophomores in the Laboratory."
24. B. H. Newberry, "Truth Telling in Subjects with Information About Experiments: Who Is Being Deceived?" *Journal of Personality and Social Psychology* 25 (1973): 369–374.
25. R. Hertwig and A. Ortmann, "Deception in Experiments: Revisiting the Arguments in Its Defense," *Ethics & Behavior* 18 (2008): 59–92.
26. C. D. Kam, "Implicit Attitudes, Explicit Choices: When Subliminal Priming Predicts Candidate Preference," *Political Behavior* 29 (2007): 343–367.
27. Basil, "The Use of Student Samples."

For Review-No Commercial Use(2023)

For Review-No Commercial Use(2023)

28. Deborah Nazarian and Joshua M. Smyth, "Context Moderates the Effects of an Expressive Writing Intervention: A Randomized Two-Study Replication and Extension," *Journal of Social & Clinical Psychology* 29, no. 8 (2010): 903–929.
29. Basil, "The Use of Student Samples."
30. Ibid.; Sears, "College Sophomores in the Laboratory."
31. Basil, "The Use of Student Samples."
32. Ibid.
33. Ibid.
34. Renita Coleman and H. Denis Wu, *Image and Emotion in Voter Decisions: The Affect Agenda* (New York: Lexington Books, 2015), chapter 9.
35. W. A. Cunningham, T. Anderson, and J. Murphy, "Are Students Real People?" *The Journal of Business* 47 (1974): 399–409.
36. Basil, "The Use of Student Samples."
37. Ibid.
38. Yair Levy, Timothy J. Ellis, and Eli Cohen, "A Guide for Novice Researchers on Experimental and Quasi-Experimental Studies in Information Systems Research," *Interdisciplinary Journal of Information, Knowledge and Management* 6 (2011): 151–161.
39. J. W. Lucas, "Theory-Testing, Generalization, and the Problem of External Validity," *Sociological Theory* 21 (2003): 236–253; J. G. Lynch Jr., "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research* 9 (1982): 225–239; John G. Lynch Jr., "The Role of External Validity in Theoretical Research," *Journal of Consumer Research* 10 (1983): 109–111.
40. Peterson, "On the Use of College Students."
41. Paul H. P. Hanel and Katia C. Vione, "Do Student Samples Provide an Accurate Estimate of the General Public?" *PLoS ONE* 11, no. 12 (2016): 1–10.
42. Coleman and Wu, *Image and Emotion in Voter Decisions.*
43. C. D. Kam, J. R. Wilking, and E. J. Zechmeister, "Beyond the 'Narrow Data Base': Another Convenience Sample for Experimental Research," *Political Behavior* 29 (2007): 415–440; P. J. Henry, "College Sphomores in the Laboratory Redux: Influences of a Narrow Databse on Social Psychology's View of the Nature of Prejudice," *Psychological Inquiry* 19, no. 2 (2008): 49–71.
44. M. H. Birnbaum, ed., *Psychological Experiments on the Internet* (San Diego, CA: Academic Press, 2000); M. H. Birnbaum, "Human Research and Data Collection Via the Internet," *Annual Review of Psychology* 55, no. 1 (2004): 803–832; Henry, "College Sphomores in the Laboratory Redux"; U. D. Reips, "The Web Experiment Method: Advantages, Disadvantages and Solutions," in *Psychological Experiments on the Internet*, ed. M. H. Birnbaum (San Diego, CA: Academic Press, 2000); M. P. Wattenberg, *Is Voting for Young People?* (Upper Saddle River, NJ: Pearson, 2011).
45. Erin C. Cassese et al., "Socially Mediated Internet Surveys: Recruiting Participants for Online Experiments," *PS: Political Science & Politics* 46, no. 4 (2013): 1–10.
46. Marriott, "Using Student Subjects in Experimental Research."
47. Cassese et al., "Socially Mediated Internet Surveys."
48. Ibid.
49. E. Diener, M. Diener, and C. Diener, "Factors Predicting the Subjective Well-Being of Nations," *Journal of Personality and Social Psychology* 69 (1995): 851–864; R. Fischer and S. Schwartz, "Whence Differences in Value Priorities? Individual, Cultural, or Artifactual Sources," *Journal of Cross-Cultural Psychology* 42 (2011): 1127–1144.
50. F. R. Kardes, "In Defense of Experimental Consumer Psychology," *Journal of Consumer Psychology* 5 (1996): 279–296; Lucas, "Theory-Testing, Generalization."
51. Basil, "The Use of Student Samples"; Lang, "The Logic of Using Inferential Statistics."
52. J. Alm and M. McKee, "Extending the Lessons of Laboratory Experiments on Tax Compliance to Managerial and Decision Economics," *Managerial and Decision Economics* 19 (1998): 259–275.
53. D. Bello et al., "From the Editors: Student Samples in International Business Research," *Journal of International Business Studies* 40 (2009): 361–364; D. G. Mook, "In Defense of External Invalidity," *American Psychologist* 38 (1983): 379–387; R. E. Pernice et al., "On Use of Student Samples for Scale Construction," *Psychological Reports* 102 (2008): 459–464.
54. Peterson, "On the Use of College Students"; Robert A. Peterson and Dwight R. Merunka, "Convenience Samples of College Students and Research Reproducibility," *Journal of Business Research* 67, no. 5 (2014): 1035–1041.

For Review-No Commercial Use(2023)

55. A. Bardi et al., "Value Stability and Change During Self-Chosen Life Transitions: Self-Selection Versus Socialization Effects," *Journal of Personality and Social Psychology* 106 (2014): 131–147.
56. M. Host, B. Regnell, and C. Wohlin, "Using Students as Subjects—A Comparative Study of Students and Professionals in Lead-Time Impact Assessment," *Empirical Software Engineering* 5 (2000): 201–214; G. Liyanarachchi, "Feasibility of Using Student Subjects in Accounting Experiments: A Review," *Pacific Accounting Review* 19, no. 1 (2007): 47–67; Marriott, "Using Student Subjects in Experimental Research."
57. Marriott, "Using Student Subjects in Experimental Research."
58. C. M. Megehee, "Advertising Time Expansion, Compression, and Cognitive Processing Influences on Consumer Acceptance of Message and Brand," *Journal of Business Research* 62, no. 4 (2009): 420–431.
59. Basil, "The Use of Student Samples."
60. For example, see: S. Hazari, C. Brown, and R. Rutledge, "Investigating Marketing Students' Perceptions of Active Learning and Social Collaboration in Blogs," *Journal of Education for Business* 88 (2013): 101–108; Host, Regnell, and Wohlin, "Using Students as Subjects"; Jin Huang et al., "Individual Development Accounts and Homeownership Among Low-Income Adults with Disabilities: Evidence from a Randomized Experiment," *Journal of Applied Social Science* 10, no. 1 (2016): 55–66; Yanna Krupnikov and Adam Seth Levine, "Cross-Sample Comparisons and External Validity," *Journal of Experimental Political Science* 1, no. 1 (2014): 59–80; Liyanarachchi, "Feasibility of Using Student Subjects"; G. Liyanarachchi and M. Milne, "Comparing the Investment Decisions of Accounting Practitioners and Students: An Empirical Study on the Adequacy of Student Surrogates," *Accounting Forum* 29 (2005): 121–135; D. W. Schanzenbach, "What Have Researchers Learned from Project Star?" in *Brookings Papers on Education Policy* (2007): 205–228; Valmi D. Sousa, Jaclene A. Zauszniewski, and Carol M. Musil, "How to Determine Whether a Convenience Sample Represents the Population 1," *Applied Nursing Research* 17, no. 2 (2004): 130–133; T. Waller, C. Lampman, and G. Lupfer-Johnson, "Assessing Bias against Overweight Individuals Among Nursing and Psychology Students: An Implicit Association Test," *Journal of Clinical Nursing* 21 (2012): 3504–3512; James N. Druckman and C. D. Kam, "Students as Experimental Participants: A Defense of the "Narrow Data Base," in *Cambridge Handbook of Experimental Political Science*, ed. James N. Druckman, et al. (New York: CUP, 2011), 41–57.
61. A. Abdel-Khalik, "On the Efficiency of Subject Surrogation in Accounting Research," *The Accounting Review* 49 (1974): 743–750; D. Bean and J. D'Aquila, "Accounting Students as Surrogates for Accounting Professionals When Studying Ethical Dilemmas: A Cautionary Note," *Teaching Business Ethics* 7 (2003): 187–204; R. Copeland, A. Francia, and R. Strawser, "Students as Subjects in Behavioral Business Research," *The Accounting Review* 48 (1973) 365–372; M. T. Gordon, A. Slade, and N. Schmitt, "The 'Science of the Sophomore' Revisited: From Conjecture to Empiricism," *Academy of Management Review* 11 (1986): 191–207.
62. Sousa, Zauszniewski, and Musil, "How to Determine Whether a Convenience Sample Represents the Population 1."
63. W. G. Cochran, *Sampling Techniques,* 3rd ed. (New York: Wiley, 1977).
64. Renita Coleman, Esther Thorson, and Lee Wilkins, "Testing the Impact of Public Health Framing and Rich Sourcing in Health News Stories," *Journal of Health Communication* 16, no. 9 (2011): 941–954; Rebecca S. McEntee, Renita Coleman, and Carolyn Yaschur, "Comparing the Effects of Vivid Writing and Photographs on Moral Judgment in Public Relations," *Journalism and Mass Communication Quarterly* 94 no. 4 (2017): 1011–1030; Aimee Meader et al., "Ethics in the Digital Age: A Comparison of the Effects of Moving Images and Photographs on Moral Judgment," *Journal of Media Ethics* 30, no. 4 (2015): 234–251.
65. Nazarian and Smyth, "Context Moderates the Effects."
66. Krupnikov and Levine, "Cross-Sample Comparisons."
67. Peterson, "On the Use of College Students"; Peterson and Merunka, "Convenience Samples of College Students."
68. R. M. Lindsay and A. S. C. Ehrenberg, "The Design of Replicated Studies," *American Statistician* 47 (1993): 236.
69. Peterson and Merunka, "Convenience Samples of College Students."

70. Basil, "The Use of Student Samples."
71. Hanel and Vione, "Do Student Samples Provide an Accurate Estimate of the General Public?"
72. Christine E. Meltzer, Thorsten Naab, and Gregor Daschmann, "All Student Samples Differ: On Participant Selection in Communication Science," *Communication Methods and Measures* 6 (2012): 251–262.
73. Coleman, "The Effect of Visuals."
74. Claire A. Dunlop and Claudio M. Radaelli, "Teaching Regulatory Humility: Experimenting with Student Practitioners," *Politics* 36, no. 1 (2016): 79–94.
75. Basil, "The Use of Student Samples."
76. T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Chicago: Rand-McNally, 1979).
77. Peterson, "On the Use of College Students."
78. Marriott, "Using Student Subjects in Experimental Research."
79. Peterson, "On the Use of College Students"; Peterson and Merunka, "Convenience Samples of College Students."
80. Marriott, "Using Student Subjects in Experimental Research."
81. Sousa, Zauszniewski, and Musil, "How to Determine Whether a Convenience Sample Represents the Population 1."
82. Winter Mason and Siddharth Suri, "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavioral Research* 44 (2012): 1–23.
83. Srinivas Rao and Amanda Michel, "ProPublica's Guide to Mechanical Turk," *ProPublica* (Oct. 15, 2010).
84. Bassam Tariq, "Turking for a Living," *New Yorker* (March 20, 2015).
85. Mason and Suri, "Conducting Behavioral Research."
86. Christoph Bartneck et al., "Comparing the Similarity of Responses Received From Studies in Amazon's Mechanical Turk to Studies Conducted Online and With Direct Recruitment," *PLoS ONE* 10, no. 4 (2015): 1–23.
87. http://en.wikipedia.org/wiki/The_Turk. Accessed June 24, 2017.
88. http://en.wikipedia.org/wiki/The_Turk. Accessed June 24, 2017.
89. http://en.wikipedia.org/wiki/The_Turk; Bartneck et al., "Comparing the Similarity of Responses."
90. J. Ross et al., "Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk" (paper presented at the ACM, 2010).
91. Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.Com's Mechanical Turk," *Political Analysis* 20, no. 3 (2012): 351–368; Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis, "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making* 5, no. 5 (2010); Ross et al., "Who Are the Crowdworkers?"
92. Berinsky, Huber, and Lenz, "Evaluating Online Labor Markets."
93. Michael Buhrmester, Tracy Kwang, and Sam Gosling, "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?" *Perspectives on Psychological Science* 6, no. 1 (2011): 3–5; R. Kosara and C. Ziemkiewicz, "Do Mechanical Turks Dream of Square Pie Charts?" *Proceedings BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization (BELIV)* 10 (2010): 373–382.
94. Connor Huff and Dustin Tingley, "'Who Are These People?' Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents," *Research and Politics* July-September (2015): 1–12.
95. Krupnikov and Levine, "Cross-Sample Comparisons."
96. Thomas Bernauer, Robert Gampfer, and Aya Kachi, "European Unilateralism and Involuntary Burden-Sharing in Global Climate Politics: A Public Opinion Perspective from the Other Side," *European Union Politics* 15, no. 1 (2014): 132–151; Robert Gampfer, Thomas Bernauer, and Aya Kachi, "Obtaining Public Support for North-South Climate Funding: Evidence from Conjoint Experiments in Donor Countries," *Global Environmental Change* 29 (2014): 118–126.
97. Huff and Tingley, "'Who Are These People?'"
98. Bartneck et al., "Comparing the Similarity of Responses," 1/23.
99. Huff and Tingley, "'Who Are These People?'" 8.
100. Kevin E. Levay, Jeremy Freeze, and James N. Druckman, "The Demographic and Political Composition of Mechanical Turk Samples," *SAGE Open* January-March (2016): 1–17
101. Ibid., 10.

For Review-No Commercial Use(2023)

102. Berinsky, Huber, and Lenz, "Evaluating Online Labor Markets."
103. Ibid.; Buhrmester, Kwang, and Gosling, "Amazon's Mechanical Turk"; Huff and Tingley, "'Who Are These People?'"; Panagiotis G. Ipeirotis, "Demographics of Mechanical Turk," *CeDER Working Paper No. 10-10* New York University, no. Retrieved from http://www.ipeirotis.com/wp-content/uploads/2012/02/CeDER-10-01.pdf (2010); Krupnikov and Levine, "Cross-Sample Comparisons"; Kevin J. Mullinix et al., "The Generalizability of Survey Experiments," 2 (2015): 109–138.
104. Berinsky, Huber, and Lenz, "Evaluating Online Labor Markets"; Krupnikov and Levine, "Cross-Sample Comparisons"; Mullinix et al., "The Generalizability of Survey Experiments."
105. Birnbaum, *Psychological Experiments on the Internet*; Paolacci, Chandler, and Ipeirotis, "Running Experiments on Amazon Mechanical Turk."
106. Buhrmester, Kwang, and Gosling, "Amazon's Mechanical Turk."
107. Bartneck et al., "Comparing the Similarity of Responses," 5/23.
108. Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis, "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research," *PloS One* 8, no. 3 (2013): 1–18.
109. Krupnikov and Levine, "Cross-Sample Comparisons."
110. Ibid.; Levay, Freeze, and Druckman, "The Demographic and Political Composition."
111. Krupnikov and Levine, "Cross-Sample Comparisons."
112. Ibid.
113. Emily Thorson, "Belief Echoes: The Persistent Effects of Corrected Misinformation," *Political Communication* 33, no. 3 (2016). 460–480.
114. Huff and Tingley, "'Who Are These People?'"
115. Ellen Cushing, "Amazon Mechanical Turk: The Digital Sweatshop," *Utne Reader* January-February, http://www.utne.com/science-and-technology/amazon-mechanical-turk-zm0z13jfzlin (2013); John Horton and Lydia Chilton, "The Labor Economics of Paid Crowdsourcing," in Proceedings of the 11th ACM Conference on Electronic Commerce (2010).
116. Bartneck et al., "Comparing the Similarity of Responses."
117. Buhrmester, Kwang, and Gosling, "Amazon's Mechanical Turk."
118. Travis Simcox and Julie Fiez, "Collecting Response Times Using Amazon Mechanical Turk and Adobe Flash," *Behavioral Research* 46 (2014): 95–111.
119. Scott Clifford and Jennifer Jerit, "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies," *Journal of Experimental Political Science* 1, no. 2 (2014): 120–131.
120. Berinsky, Huber, and Lenz, "Evaluating Online Labor Markets."
121. Bartneck et al., "Comparing the Similarity of Responses"; Berinsky, Huber, and Lenz, "Evaluating Online Labor Markets"; A. J. Berinsky, M. F. Margolia, and M. W. Sances, "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys," *American Journal of Political Science* 58, no. 3 (2014): 739–753; Huff and Tingley, "'Who Are These People?'"
122. D. Rand, "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments," *Journal of Theoretical Biology* 299 (2012): 172–179.
123. J. Heer and M. Bostock, "Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design" (paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, 2010).
124. Buhrmester, Kwang, and Gosling, "Amazon's Mechanical Turk"; Mason and Suri, "Conducting Behavioral Research."
125. Paolacci, Chandler, and Ipeirotis, "Running Experiments on Amazon Mechanical Turk."
126. Buhrmester, Kwang, and Gosling, "Amazon's Mechanical Turk"; Heer and Bostock, "Crowdsourcing Graphical Perception"; D. Kelly, D. Harper, and B. Landau, "Questionnaire Mode Effects in Interactive Information Retrieval Experiments," *Information Processing and Management* 44 (2008): 122–141.
127. Bartneck et al., "Comparing the Similarity of Responses"; Simcox and Fiez, "Collecting Response Times."
128. Paolacci, Chandler, and Ipeirotis, "Running Experiments on Amazon Mechanical Turk."
129. M. Smith and B. Leigh, "Virtual Subjects: Using the Internet as an Alternative Source of Subjects and Research Environment," *Behavior Research Methods* 29 (1997): 496–505.

For Review-No Commercial Use(2023)

130. Marriott, "Using Student Subjects in Experimental Research."
131. Gregg R. Murray et al., "Convenient Yet Not a Convenience Sample: Jury Pools as Experimental Subject Pools," *Social Science Research* 42, no. 1 (2013): 246–253.
132. Alan Gerber et al., "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee," *Journal of Experimental Political Science* 1, no. 1 (2014): 81–98.
133. Alan S. Gerber et al., "Reporting Balance Tables, Response Rates and Manipulation Checks in Experimental Research: A Reply from the Committee That Prepared the Reporting Guidelines," *Journal of Experimental Political Science* 2, no. 2 (2016): 216–229.
134. Diana C. Mutz and Robin Pemantle, "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods," *Journal of Experimental Political Science* 2, no. 2 (2016): 192–215.
135. Bartneck et al., "Comparing the Similarity of Responses."
136. Cassese et al., "Socially Mediated Internet Surveys."
137. Ibid.
138. Ibid., 777.
139. Kam, "Implicit Attitudes, Explicit Choices."
140. Murray et al., "Convenient Yet Not a Convenience Sample."
141. For an example of a study that incorporated experimental conditions in the British Election Study 2012, see Rosie Campbell and Philip Cowley, "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment," *Political Studies* 62, no. 4 (2013): 745–765; Murray et al., "Convenient Yet Not a Convenience Sample."
142. David Niven, "The Other Side of Optimism: High Expectations and the Rejection of Status Quo Politics," *Political Behavior* 22, no. 1 (2000): 71–88.
143. Lee Sigelman, Carol K. Sigelman, and Barbara J. Walkosz, "The Public and the Paradox of Leadership: An Experimental Analysis," *American Political Science* 36 (1992): 366–385.
144. Nayda Terkildsen, "When White Voters Evaluate Black Candidates: The Processing Implications of Candidate Skin Color, Prejudice, and Self-Monitoring," *American Journal of Political Science* 37 (1993): 1032–1053.
145. Christopher J. Lewis, Dona-Gene Mitchell, and Cynthia R. Rugeley, "Courting Public Opinion: Utilizing Jury Pools in Experimental Research," in *Political Methodology Conference* (Tallahassee, FL 2005).
146. Ibid.
147. Murray et al., "Convenient Yet Not a Convenience Sample," 252.
148. McEntee, Coleman, and Yaschur, "Comparing the Effects of Vivid Writing."
149. Mitchell Hoffman and John Morgan, "Who's Naughty? Who's Nice? Experiments on Whether Pro-Social Workers Are Selected out of Cutthroat Business Environments," *Journal of Economic Behavior & Organization* 109 (2015): 173–187.
150. Kim and Ryzin, "Impact of Government Funding"; Murray et al., "Convenient Yet Not a Convenience Sample."
151. Gary W. Ritter and Marc J. Holley, "Lessons for Conducting Random Assignment in Schools," *Journal of Children's Services* 3, no. 2 (2008): 28–39.
152. Ibid.
153. Willem Wetzels et al., "Impact of Prepaid Incentives in Face-to-Face Surveys: A Large-Scale Experiment with Postage Stamps," *International Journal of Public Opinion Research* 20, no. 4 (2008): 507–516.
154. Dunlop and Radaelli, "Teaching Regulatory Humility."
155. S. Mayr et al., "A Short Tutorial of Gpower," *Tutorials in Quantitative Methods for Psychology* 3, no. 2 (2007): 51–59.
156. Ibid.
157. Ibid.
158. Ibid.
159. For example, see J. Cohen, "Things I Have Learned (So Far)," *American Psychologist* 45 (1990): 1304–1312; Jacob Cohen, "The Earth Is Round (P < .05)," *American Psychologist* 49, no. 12 (1994): 997–1003.
160. L. J. Cronbach et al., *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles* (New York: Wiley, 1972); G. A. Marcoulides, "Maximizing Power in Generalizability Studies under Budget Constraints," *Journal of Educational Statistics* 18, no. 2 (1993): 197–206.
161. Wilson Van Voorhis and Morgan, "Understanding Power and Rules of Thumb."
162. Barbara G. Tabachnick and Linda S. Fidell, *Using Multivariate Statistics* (New York: HarperCollins, 1996);

For Review-No Commercial Use(2023)

Esther Thorson, Robert H. Wicks, and Glenn Leshner, "Experimental Methodology in Journalism and Mass Communication Research," *Journalism and Mass Communication Quarterly* 89, no. 1 (2012): 112–124.
163. J. Descôteaux, "Statistical Power: An Historical Introduction," *Tutorials in Quantitative Methods for Psychology* 3, no. 2 (2007): 28–34; Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
164. Rebecca B. Morton and Joshua A. Tucker, "Experiments, Journals, and Ethics," *Journal of Experimental Political Science* 1, no. 2 (2015): 99–103.
165. Chia-Hu Chang et al., "Virtual Spotlighted Advertising for Tennis Videos," *Journal of Visual Communication & Image Representation* 21, no. 7 (2010): 595–612; Justin W. Owens, Evan M. Palmer, and Barbara S. Chaparro, "The Pervasiveness of Text Advertising Blindness," *Journal of Usability Studies* 9, no. 2 (2014): 51–69.
166. James R. Rest, *Development in Judging Moral Issues* (Minneapolis, MN: University of Minnesota Press, 1979).
167. Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
168. Leland Wilkinson and Task Force on Statistical Inference, American Psychological Association Science Directorate, "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist* 54, no. 8 (1999): 594–604.
169. Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
170. D. R. Atkinson, M. J. Furlong, and B. E. Wampold, "Statistical Significance, Reviewer Evaluations, and the Scientific Process: Is There a (Statistically) Significant Relationship?" *Journal of Counseling Psychology* 29 (1982): 189–194.
171. Graham N. Dixon, "Negative Affect as a Mechanism of Exemplification Effects," *Communication Research* 43, no. 6 (2015): 768.
172. H. Denis Wu and Renita Coleman, "The Affective Effect on Political Judgment: Comparing the Influences of Candidate Attributes and Issue Congruence," *Journalism and Mass Communication Quarterly* 91, no. 3 (2014), 530–543.
173. Renita Coleman et al., "Public Life and the Internet: If You Build a Better Website, Will Citizens Become Engaged?" *New Media & Society* 10, no. 2 (2008): 179–201.
174. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. (Hillsdale, NJ: Erlbaum, 1988).
175. Russell V. Lenth, "Java Applets for Power and Sample Size," Division of Mathematical Sciences, the College of Liberal Arts of the University of Iowa, https://homepage.divms.uiowa.edu/~rlenth/Power/.
176. Erik Westlund and Elizabeth A. Stuart, "The Nonuse, Misuse, and Proper Use of Pilot Studies in Experimental Evaluation Research," *American Journal of Evaluation* 38, no. 2 (2017): 246–261.
177. Descôteaux, "Statistical Power."
178. Wilkinson and Task Force on Statistical Inference, American Psychological Association Science Directorate, "Statistical Methods in Psychology Journals."
179. S. B. Green, "How Many Subjects Does It Take to Do a Regression Analysis?" *Multivariate Behavioral Research* 26 (1991): 499–510.
180. R. J. Harris, *A Primer of Multivariate Statistics,* 2nd ed. (New York: Academic Press, 1985).
181. J. Cohen, and P. Cohen, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Hillsdale, NJ: Erlbaum, 1975).
182. D. C. Howell, *Statistical Methods for Psychology*, 4th ed. (Belmont, CA: Wadsworth, 1997).
183. Tabachnick and Fidell, *Using Multivariate Statistics.*
184. E. J. Pedhazur and L. P. Schmelkin, *Measurement, Design, and Analysis: An Integrated Approach* (Hillsdale, NJ: Erlbaum, 1991).
185. A. L. Comrey and H. B. Lee, *A First Course in Factor Analysis,* 2nd ed. (Hillsdale, NJ: Erlbaum, 1992).
186. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed.
187. A. Aron and E. N. Aron, *Statistics for Psychology,* 2nd ed. (Upper Saddle River, NJ: Prentice Hall, 1999).
188. Also see: F. Faul et al., "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences," *Behavior Research Methods* 39, no. 2 (2007): 175–191; Mayr et al., "A Short Tutorial of Gpower."
189. Coleman, "Journalists' Moral Judgment About Children."
190. Coleman, "The Effect of Visuals."
191. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed.
192. Descôteaux, "Statistical Power."
193. Sarah V. Bentley, Katharine H. Greenaway, and S. Alexander Haslam, "An Online Paradigm for

For Review-No Commercial Use(2023)

Exploring the Self-Reference Effect," *PLoS ONE* 12, no. 5 (2017): 1–21.

194. Erica A. Hornstein and Naomi I. Eisenberger, "Unpacking the Buffering Effect of Social Support Figures: Social Support Attenuates Fear Acquisition," *PLoS ONE* 12, no. 5 (2017): 1–9.
195. Satoris S. Culbertson, William S. Weyhrauch, and Christopher J. Waples, "Behavioral Cues as "Indicators of Deception in Structured Employment Interviews," *International Journal of Selection & Assessment* 24, no. 2 (2016): 119–131.
196. John M. Hoenig and Dennis M. Heisey, "The Abuse of Power," *The American Statistician* 55, no. 1 (2001): 19–24; Daniel J. O'Keefe, "Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses," *Communication Methods and Measures* 1, no. 4 (2007): 291–299; J. Descôteaux, "Editor's Note: The Uncorrupted Statistical Power," *Tutorials in Quantitative Methods for Psychology* 3, no. 2 (2007): 26–27.
197. Hoenig and Heisey, "The Abuse of Power."
198. Ibid.; O'Keefe, "Post Hoc Power, Observed Power."
199. Hoenig and Heisey, "The Abuse of Power"; O'Keefe, "Post Hoc Power, Observed Power."
200. Hoenig and Heisey, "The Abuse of Power."
201. O'Keefe, "Post Hoc Power, Observed Power."
202. Ibid.
203. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed.; Hoenig and Heisey, "The Abuse of Power"; O'Keefe, "Post Hoc Power, Observed Power."

For Review-No Commercial Use(2023)

9

STIMULI AND MANIPULATION CHECKS

Look for a large object rather than a small one.[1]

—R. Barker Bausell

For Review Only Commercial Use(2023)

LEARNING OBJECTIVES

- Explain how stimuli represent categories of theoretical interest.
- Create realistic stimuli.
- Judge variables that need to be controlled in the stimuli versus those that should be controlled statistically.
- Design stimuli that maximize comparisons and employ message variance.
- Define fixed and random factors.
- Produce a report that details the stimuli and manipulation check of an experiment.

The heart of any experiment is the stimulus. This is also the fun part. Experimentalists seldom define the term **stimulus**, instead relying on the dictionary definition of it as something that causes a response or reaction.[2] For example, Freud called a stimulus a motivating force, mirroring the dictionary definition.[3] The most common usage of the term is likely in the context of the economy—discussion of a "stimulus package" is

STUDY SPOTLIGHT 9.1

Creating Realistic Stimuli

SAGE Journal Article: study.sagepub.com/coleman

From: Chen, Gina Masullo. 2013. "Losing Face on Social Media." Communication Research 42 (6): 819–838. doi:10.1177/0093650213510937.

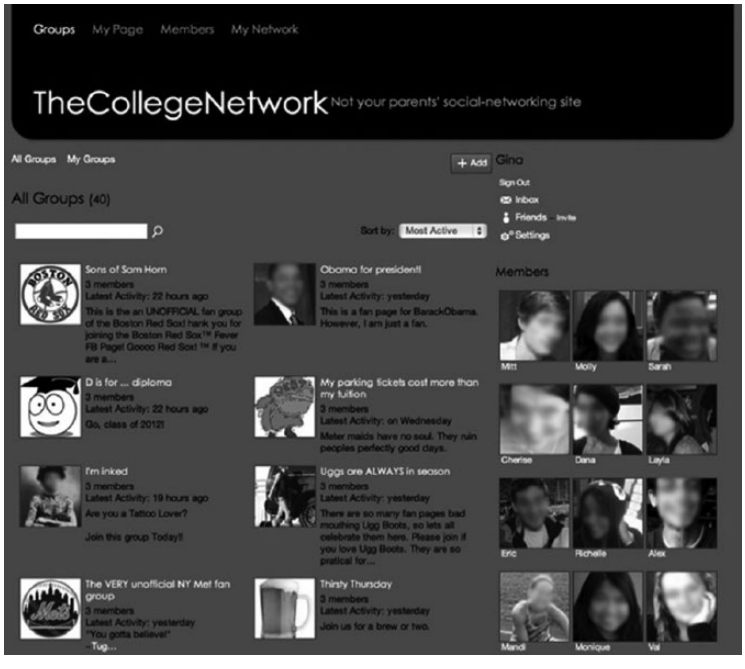
In this study, the author goes to great lengths to create realistic stimuli and conduct manipulation checks in order to create a fictitious but believable social networking site for college students. She first began with a focus group of seven college students and used a procedure that asked them to think out loud (see chapter 10 for more on this) while brainstorming ideas for the topic of the site (e.g., "Leggings Aren't Pants," "How Do They Expect Me to Learn at 8 a.m. When I'm Still Drunk"), questions on it (e.g., "Top 5 songs on my iPod are . . ." "On the weekend, you're most likely to find me . . ."), and the types of messages they might receive when trying to join such a site.

She used a customizable online platform called Ning to create the fictitious social networking site called "The College Network." Subjects were told they were beta testing a new social media site for college students.

For this, she adapted a procedure advocated by Graesser (1981) to create stimuli for experiments. Information about this book can be found in the "Suggested Readings" section at the end of this chapter.

FIGURE 9.1 ⬢ THE COLLEGE NETWORK

For Review-No Commercial Use(2023)



Gina Masullo Chen

pervasive when the economy is sluggish. When politicians and economists talk about a stimulus, they mean something that will increase employment and encourage spending. They are interested in doing something that will cause the economy to react positively. The term is similar for social science experiments—a stimulus is the vehicle that an experimenter uses to deliver the manipulation hypothesized to cause some effect. And as the advice from Bausell in the opening quote says, a "large object"—in this case, a big or strong stimulus—is better suited to create a response than is a small one.

Stimuli (the plural form of stimulus) for an experiment can be in whatever form best delivers the manipulation, treatment, or intervention to the subjects. For Ivan Pavlov, the stimulus was a ringing bell. Stanley Milgram built an "electric shock generator" as a stimulus, which was nothing more than a box with switches and labels. Philip Zimbardo created an entire mock prison as his stimulus. In many different disciplines in the social sciences, researchers commonly use messages as a principal source of stimuli[4]—for example, campaign websites in political studies, advertisements for marketing research, clips of movies to induce mood in a psychology study. Stimuli are often conceived of in terms of their physical features,[5] such as the functionality or organization of a website, which makes up the concept of interest.[6] Another form of stimulus is as instructions—for example, telling subjects to list what worries them about an issue such as immigration.[7] Stimuli can range from the very simple, such as text on a piece of paper or computer screen, to the complex, such as an entire fictitious TV news show or website. Many are somewhere in between—posters, web pages, press releases, tweets, social media posts, and the like. One political study sent subjects text messages as the stimuli.[8] For some disciplines that assess the effectiveness of programs or interventions, the stimuli are those programs or interventions, such as a computer-assisted learning curriculum or a training seminar given to police to teach them how to better handle people with mental illness, for example. In these cases, the stimuli already exist, and researchers may only need to create something else for the control group to receive.

For Review-No Commercial Use(2023)

Creating high-quality stimuli is key to a successful experiment, so this chapter is devoted to advice and suggestions to achieve this.

EXAMPLES OF STIMULI

Probably the best way to illustrate a stimulus in a social science experiment is with examples:

- In a political science study aimed at determining which politicians make it into the news and why, the stimuli were fictional press releases about parliamentarians.[9]

- In a marketing experiment designed to see what messages were most effective at getting people to cook on fuel-efficient stoves instead of using fires in Uganda, the experimenters created posters touting the new stoves.[10]
- In a study of a public relations crisis, the voice pitch of a spokesperson was manipulated with radio broadcasts of a fictitious press conference.[11]
- In an education study, students were taught on a computer instead of by a teacher; the computer-assisted curriculum was the stimulus.[12]

In all these examples, the stimuli are designed to represent the categories of theoretical interest to the study. They are the vehicle through which the experimenters manipulate the levels of the factors hypothesized to cause some effect or outcome. Specifically:

- In the political science study, the press releases were written to represent characteristics hypothesized to make politicians newsworthy, such as which party they belonged to and the issue they specialized in.[13] These represent the levels of the factors that were manipulated, or changed, in the stimulus press releases.
- In the marketing experiment for stoves, the posters were the stimuli that delivered the manipulated messages representing the levels of the factor—for example, one on health benefits and one on saving time and money.[14]
- In the PR crisis study, the radio broadcasts of a press conference were the stimuli created by having a student pretending to be an organization's spokesperson record a statement, which was then manipulated to create one high-pitched version and one low-pitched version; the pitch represented the levels of the factor being studied.[15]
- In the education experiment, the computer-assisted lessons were the stimuli that manipulated the concepts in a theory of language learning that was being tested by changing how visual and auditory information was presented, and with different question-and-answer strategies, for example.[16]

For Review-No Commercial Use(2023)

In journalism, stories are often created to look like they have appeared in a newspaper or on a website. For example, a study of the effects of speed versus accuracy on readers' perceptions of news organizations' credibility created stimuli that consisted of stories in blog-roll format that showed inaccurate information being corrected as the story updated.[17]

In this study of the effects of speed and accuracy on readers' perceptions of credibility of news organizations, the researcher created stimuli designed to look like a news organization's blog roll. These were presented to subjects on a computer screen the way they would

see them if they were real. She imitated actual news organization's blog rolls to create the stimuli. Figures 9.2a and 9.2b show an example of a story in the fast condition, with inaccuracies, and in the slow condition, with no inaccuracies. In the fast condition, the manipulation is achieved with the term "CORRECTION."

FIGURE 9.2A ⬢ FAST CONDITION

Lee, Angela Min-Chia. 2014. How Fast Is Too Fast? Examining the Impact of Speed-Driven Journalism on News Production and Audience Reception (unpublished dissertation, The University of Texas at Austin).

Dozens Dead in San Antonio Flooding, Landslides

Live Text As It Happened

By NICHOLAS KULISH

10:02 a.m.: Torrential rains began on Sunday night and continued into Monday morning, causing power failures and cutting off roadways near The Far Northwest.

9:40 a.m.: Mayor Martinez de Vera said burials for the dead and treatment for the injured are being organized.

9:28 a.m.: "Many houses in the suburbs were destroyed by the rain," Greenberg said. "Many of the victims are children, and the houses collapsed on them."

For Review-No Commercial Use(2023)

9:23 a.m.: Police spokesman, Jon Hermes, said that hundreds more had been displaced because of damage to their homes. Willy Greenberg, spokesman for Mayor Art Martinez de Vera, said the damage is extensive.

9:17 a.m.: *CORRECTION*—At least 60, not 100, people were killed.

9:12 a.m.: At least 100 people were killed and dozens injured in San Antonio, after heavy rains caused flooding and landslides in Bexar County, officials said.

FIGURE 9.2B ⬢ SLOW CONDITION

Dozens Dead in San Antonio Flooding, Landslides

By NICHOLAS KULISH

At least 60 people were killed and dozens injured in San Antonio after heavy rains caused flooding and landslides in Bexar County, officials said.

Police spokesman Jon Hermes said hundreds more had been displaced because of damage to their homes. Willy Greenberg, spokesman for Mayor Art Martinez de Vera, said the damage is extensive.

"Many houses in the suburbs were destroyed by the rain," Greenberg said. "Many of the victims are children, and the houses collapsed on them."

Martinez de Vera said burials for the dead and treatment for the injured are being organized. Torrential rains began on Sunday night and continued into Monday morning, causing power failures and cutting off roadways near The Far Northwest.

Quite a lot of work is involved in creating realistic-looking posters, radio broadcasts, ads, and websites; however, not all stimuli need to be this complex. In a business study on corporate social responsibility, the researchers manipulated how socially responsible a business was by describing it as either nonprofit or for-profit in text scenarios that participants read; the scenarios were the stimuli.[18] Text on paper or on computer screens is among the simplest to create, and many experiments use these as stimuli. For example, in a political science study where subjects were given information about fictional candidates' positions on issues and then asked to cast a vote, the information was presented as text on a computer screen rather than being billed as coming from a candidate's website or a news source, eliminating the need to mock up a website or newspaper editorial.[19] In another experiment, subjects were given an essay to read that primed their feelings of disgust or concerns about harm, and then asked to make a moral judgment about gay adoption.[20] The essay was not presented as coming from an editorial in a magazine or newspaper, so there was no need to make it look like one. However, if an experiment calls for some message that is typically delivered via a medium such as TV commercials or Twitter, then having subjects read text that describes a TV commercial or tweet is far less realistic than creating a fictional TV commercial.

This is one of the reasons why there is so much criticism about the lack of external validity in experiments (see chapter 5 on validity). (For an online tool that creates fake social media such as Twitter, Facebook, and more, see More About box 9.2.)

For Review-No Commercial Use(2023)

Sometimes, the stimulus itself is not complicated to create but the accompanying parts of the experiment are. For example, if the manipulation is embedded in written instructions, as it was in an economics experiment where subjects performed a task and were told that a contribution would either go to an organization or not, there is no need to create stimuli such as posters or ads. The written instructions are the stimuli. However, the researchers did need to create the task that subjects would perform—decoding letters into numbers. Thus, they used software to create tables consisting of one column of letters and an adjacent column with the corresponding numbers, which was displayed to subjects on a computer screen.[24]

Stimuli do not always have to be fictitious or created by researchers; some studies use the real thing. For example, in a study designed to see if reality TV shows could encourage altruistic behavior, researchers used actual segments from reality shows.[25] The TV shows were the stimuli, selected to represent either game-style reality shows (*Survivor, Amazing Race*) or meaningful and inspiring reality shows (*Undercover Boss* and *Supernanny*). In another study, three stimuli testing the effects of Internet messages on political efficacy were an online quiz created by Minnesota Public Radio, a story from the TV

MORE ABOUT . . . BOX 9.2
Creating Realistic Social Media Stimuli

More and more information is delivered to people via social media such as Facebook and Twitter—especially to millennials, who represent the college students so often used as subjects in academic experiments. Online tools are available to help researchers create believable social media. For example, Twister, Tweeterino, and Simitator all offer tools to create fictitious Tweets, with Simitator also able to create fictitious Facebook posts and chats, SnapChat messages, and iPhone texts. PrankMeNot is another source for creating tweets, Facebook posts, and message chats. The Simitator website touts the ability to "prank your friends," but educators have also found it useful as a teaching tool in schools that block Internet access for students,[21] for instance, or to create handouts that students can relate to.[22]

One researcher who used Simitator to create believable Tweets for experiments[23] started by randomly sampling actual published tweets from the *New York Times*, *Wall Street Journal*, and *Washington Post*. He then used the tool to recreate the tweets but made them look like they came from fictitious journalists. He also simulated the way journalists cover multiple stories throughout the day for a realistic Twitter-following experience.

For a more ambitious project, creating a fully working social media site is possible with Ning, although it is not interactive. It allows the creator to add news stories, pictures, headlines, and to manipulate the number of likes and followers. Another site that allows for custom editing of web pages is CloneZone. No doubt there are others. These are free, but some others charge.

For Review-No Commercial Use(2023)

Another way to adapt real web pages for experiments is by using the "inspect element" tool in a web browser. Simply open up a web page and right click, then scroll down to the bottom of the box and select "Inspect Element." A source code pane will open up at the bottom. Click and drag to highlight the text you wish to change in the top pane, then right click and select "Inspect Element" again. The text will be highlighted in the bottom pane. Simply type over the old words and the new words will automatically appear on the website. Close the bottom pane by clicking the X on the right side, then take a screenshot and use this as a (non-interactive) web page in an experiment. You can use Command+Shift+4 on a Mac, then use the crosshair pointer to edit out the top and bottom of the browser. On a PC, use Alt+PrtScn or Snipping Tool. The changes are not permanent; the page returns to normal once you refresh it. There are tutorials online with more detail.

Find these tools at:

Ning: https://www.ning.com/

CloneZone: clonezone.link

Prank Me Not: www.prankmenot.com

Twister: http://www.classtools.net/twister/

(Continued)

(Continued)

Tweeterino: https://tweeterino.com/

Simitator: www.Simitator.com

At top is the original tweet from the New York Times. The text was used to create realistic tweets using fictitious journalists as stimuli for an experiment.

From: Boulter, Trent. 2017. "Following the Familiar: Effect of Exposure and Gender on Credibility of Journalists on Twitter." (Dissertation, The University of Texas at Austin),https://repositories.lib.utexas.edu/bitstream/handle/2152/62635/BOULTER-DISSERTATION-2017.pdf?sequence=1&isAllowed=y).

FIGURE 9.3 ⬢ TWEETS CREATED FOR STUDY

For Review-No Commercial Use(2023)

show *20/20*, and the MyBarackObama.com video.[26] Actual news photographs that have run in various magazines and newspapers can also be used as the stimuli that convey the treatment. In the educational study described earlier, a software company created the stimulus—the computer-assisted curriculum. Frequently, studies in education, social work, and other fields that test interventions or assess programs already or about to be put into practice in the field use preexisting stimuli.

ADVICE ON CREATING STIMULI

Keep It Real

When stimuli are fictitious, researchers should endeavor to make them as realistic as possible. One way to achieve this is to start with actual news articles, ads, posters, study guides, websites, or whatever, and then rewrite or edit them to reflect the manipulation. For example, in a study of how health stories were framed, the researchers used real stories and then rewrote them to replace all instances of one kind of frame with another that was being studied.[27] Researchers can also use actual stimuli, such as study guides, movies, or TV broadcasts, and edit in different wording or video to reflect the variation of the study. Actual candidates' websites can be altered to remove party identification, change the issue positions, and replace names and photos with fictitious ones. (For an example of advertisements created this way, see figure 9.4.)

For Review-No Commercial Use(2023)

Here are some examples of advertisements used as stimuli in an experiment. The researchers collected actual advertisements in order to make the stimuli as realistic as possible and then altered them to meet the requirements of the treatment and control conditions.

Another way to achieve more realistic stimuli is to have professionals (or former professionals, often graduate students) write the press releases, create websites, alter study guides, etc. Researchers may think they can create a realistic-looking newspaper clip, for example, but the finished product will have the headlines centered and in all capital letters, which real newspapers tend not to do. There are a lot of subtle things that go into creating a professional-looking product that nonprofessionals are not always aware of. Two researchers acknowledge this in a study where they created news articles, writing, "We undertook considerable effort to adjust the news articles to the common lay-out and journalistic style of day-to-day Dutch news coverage."[28] On some subconscious level, subjects are likely to recognize an amateur effort and not respond in the

FIGURE 9.4 ⬢ ADVERTISEMENTS

Source: Kim, Sunny Jung, and Jeffrey T. Hancock. 2016. "How Advertorials Deactivate Advertising Schema." *Communication Research*, 10.1177/0093650216644017.

For Review-No Commercial Use(2023)

same way as if they believed the stimuli were real. This does not always require a large outlay of funds, as often students with professional knowledge can be recruited to help with these tasks.

Control as Much as Possible

Use the literature and common sense to recognize what needs to be controlled when creating stimuli. Fictitious events, people, and places are frequently used in experiments to avoid contamination by subjects' prior knowledge about real people, places, or events. Thus, in the voice pitch study,[29] the researchers used a fake organization and a fictitious health crisis. To prevent subjects from suspecting they were fake, they set the event in a different area of the country, reasoning that subjects might think it was real and they just had not heard about it because it was not in their geographic area. This represents a form of control known as **experimental control**, when the stimuli themselves are designed to be free of confounds. Another example of this is in a study of politicians in the news that was careful to select political issues that were not clearly linked to one particular party. Steering clear of issues that are favorites of either the Republicans or Democrats avoids having subjects rate the fictional candidate based on party rather than the thing being manipulated.[30] This is an experimental control because something in the stimulus is being controlled.

In other situations, using fictitious stimuli might compromise the findings. One group of researchers reasoned that might be the case in a study of what makes politicians newsworthy, and so used real members of the Dutch parliament.[31] To control for the subjects' previous knowledge about the real politicians, the study asked subjects how often they thought each parliamentarian succeeded in making it into the news and then used that as a covariate in the analysis, removing the effects of prior knowledge from the results. This is a form of control known as **statistical control**; when it is not possible or desirable to control something in the stimuli, researchers measure it and use it as a covariate in the statistical analysis. When there are things that cannot be controlled in a manipulation—how much people already know about something, for example—then it should be measured and controlled for statistically. Be sure to collect the information on this *before* giving subjects the treatment if there is any chance that they could have learned it from the manipulation. The variables used as covariates should all be something related to the outcome variable; if the dependent variable is assessment of credibility, and gender of the source is not related to credibility, then there is probably no need to measure and control for it. Gender and race are variables that are frequently thought to interact with manipulated variables and confound the results; however, this depends on what is being studied. Consult the literature.

For Review-No Commercial Use(2023)

In summary, there are two ways to control for confounds: with experimental control, which is in the manipulation or stimuli, and with statistical control, by measuring and including it in the analysis as a covariate. For example, if race is important in how a person rates the interaction between a police officer and citizen in a video, then having half of the officers in the video be White and half Black is an experimental control. If the race of the subject is important, then measuring and including the subjects' race in the analysis is a statistical control.[i] One should always prefer experimental controls to statistical controls if possible.

Vary Only the Things Being Studied

Everything except the thing being manipulated should be kept the same in all the control and treatment versions. In a study of voice pitch,[32] that meant keeping the speed with which the statement was read and the emphasis on certain words, pauses, accent, pronunciation, and wording the same for all the different recordings. Another study used eight press releases that were all written to be the same length, had the same kind of a quote, and were all negatively framed.[33] This will often be described as something "being held constant"; for example, in the political study the negative framing was described this way: "We held

[i] I thank an anonymous reviewer for providing this clear and concise example.

the news value of conflict constant . . . with all MPs criticizing the government."[34] When things vary that are not meant to, the study is said to be confounded. This was explained in chapter 1 and can be as innocent as designing an advertisement in horizontal format for the treatment group and as a vertical for the control group, or putting one picture at the top for the treatment group and at the bottom for the control group. In these cases, there is no way to know if the horizontal or vertical format or picture placement is what actually caused any effects. It would be better to keep all ads horizontal and all pictures at the top.

Confounds can be created not just when designing stimuli but also when the experiment is conducted. For example, using two different researchers to run the study, or giving some subjects the treatment in the morning and others in the afternoon, also can be a form of confounding. Frequently, researchers are careful to hold constant the elements of the stimuli but forget about the procedural components. Any unavoidable differences can be compensated for by random assignment (see chapter 7)—for example, by randomly assigning the two researchers to run subjects in both treatment and control groups. But it is much better to avoid confounds entirely. Bausell recommends "a microscopic examination of every element in the study's design"[35] in order to uncover any possible differences, whether in the stimuli or the procedure. This involves scrutinizing every conceivable aspect of an experiment and possibly even performing a dry run or dress rehearsal, called a *pilot study*, which is a subject of chapter 11.

For Review-No Commercial Use(2023)

As with most everything else about experiments, common sense and the literature should be used as a guide in determining what is likely to be related to the outcome variable and thus should be controlled.

MAXIMIZE COMPARISONS

When creating stimuli for experiments, it is important to make the manipulation or treatment strong enough to detect a true effect. This means maximizing the comparisons between the treatment and control groups, what Bausell calls looking for large objects rather than small ones.[36] This is so important, it has even been called the first rule of experimentation.[37] Weak manipulations are a common problem that can have important consequences.[38] For example, if a study is interested in determining the effects of a newspaper's reputation on readers' assessments of credibility, the researcher would be wise to use news organizations at opposite ends of the reputation spectrum, such as the *New York Times* and the *National Enquirer*. If no effects are found with this size of a chasm in reputation, then it is highly unlikely that more subtle differences will produce effects. However, if effects *are* found between the highest- and lowest-reputation newspapers, the researcher can then go on to explore differences among newspapers with less of a reputation gap. The studies

of photographs' effects on moral judgment[39] used photographs that had won prestigious national awards, which have been shown to represent the highest level of the concept of novelty. If the study did not find effects with the strongest manipulation possible—the award-winning novel photographs—then it would be unlikely to find effects with lesser manipulations, such as photographs of the everyday variety. Award-winning photographs are not what people usually see in their daily newspapers. Because effects were found with the award-winning photographs, it would be logical to then go on to test photographs that are more commonly used in local daily newspapers, eventually testing the least novel type of photograph, such as head shots.

Stanley Milgram employed this approach in his studies of obedience to authority.[40] In the first study, he had the experimenter in the white lab coat that was ordering subjects to continue giving shocks stand in the same room, right next to the subjects. When he found effects at this close proximity, he began moving the experimenter farther and farther away from the subjects. Eventually, the experimenter was out of the room entirely. As the authority figure got farther away from the subject, obedience declined and the subjects refused to administer shock treatments.[41]

Making the manipulation that the treatment and control groups receive as different as possible is important because you like[illegible] the experiment will find a positive result, should one actually exist.[42] For example, if the study calls for manipulating the framing in a story, then make half or more of the story contain the frame being studied. This is much more than a real story would contain, but if no effects are found when so much of the story is framed a certain way, it is unlikely there would be effects when even less of the story is framed that way. Many researchers underestimate the strength of a manipulation needed to produce results,[43] so something that seems like overkill may in fact not even register with subjects.

For Review No Commercial Use (2023)

Employ Message Variance

Up to this point, this chapter has mainly been concerned with how to create variance in the stimuli based on the treatment or manipulation—that is, the levels of the independent variables of theoretical interest to the study. That is referred to as **treatment variance**.[44] But there is also another kind of variance that needs to be considered when creating stimuli—that is, message variance. **Message variance** refers to the number of messages or stimuli used to represent each treatment level.[45] For example, in studies of how the race of people affects attitudes and judgments, photographs are often used as the stimuli to manipulate the levels of the independent variable race. Rather than using one photograph of a Black person and one photograph of a White person, using three photographs of people of each race is a better strategy. This results in a

multiple-message design because it used more than one "message"—in this case a photograph—per treatment level. If only one photograph is used to represent each race, it would be a **single-message design**. When there is more than one message or stimulus to represent each level of a treatment or factor, it may also be referred to as a **repetition factor**.[46] It is important to use more than one stimulus to represent a factor, because rarely are researchers interested in a specific stimulus or message, such as one particular person, TV show, or commercial, but they are actually interested in the category, or level of the factor, that person, show, or commercial represents. For example, in the reality TV study,[47] the researchers were interested in two kinds of shows—game-style and inspiring shows. They used two shows at each level in a multiple-message design: two episodes of *Survivor* and *Amazing Race* to represent game-style, and two episodes of *Supernanny* and *Undercover Boss* to represent inspiring shows. This is important because stimuli or messages will vary in their effects, no matter how carefully researchers attempt to make them similar.[48] Multiple versions of the stimuli help ensure that effects are not due to the individual attributes of a single stimulus.

Single-message designs also are much weaker in allowing researchers to generalize beyond one specific message or stimulus—in this example, one TV show.[49] Dependable generalization requires multiple stimuli.[50] Had the reality TV researchers only used one show to represent each factor level, they only would have been able to say any effects were due to a specific episode of *Survivor* or *Undercover Boss,* not the more general categories these two shows represented. That is, they would not be able to say that a certain type of show caused effects, only that exactly the one show used had an effect. It is rarely useful to know that one particular episode of *Survivor* caused some outcome, and in any event, there may well be unknown features of that one episode that could have caused the outcome, thus confounding the results. For example, one episode may have angered or offended subjects, had a smoother transition to the commercial, or been more compatible in tone with the commercial.[51]

For Review-No Commercial Use(2023)

It is never possible to account for all confounds with a single stimulus. By using multiple stimuli to represent each level of a factor, the chances of anything other than the independent variable of interest being responsible for the outcome are reduced. As an illustration of why it is important to use more than one stimulus or message, a study of how health stories were framed used messages representing four different health issues and obtained similar effects for three of the four.[52] However, one issue—smoking—did not produce the same results. There was something about the idiosyncrasies of that particular issue or story that made people react differently to it than the others. Had the study used only this one story, it would never have been known that the frames actually worked in other contexts.

How Many Stimuli to Create?

Obviously, creating multiple stimuli is more work for the researcher. But exactly how many messages or stimuli does an experiment need? The answer is "it depends." Some studies, such as those employing latency responses where the measure is how long a subject takes to react, are unreliable unless a large number of different stimuli are used.[53] Thorson and colleagues recommend using as many as you can "reasonably expect a participant to attend to without fatigue or boredom setting in."[54] In a study of TV public service announcements (PSAs), the researchers used six stimuli for each factor level.[55] A good idea is to start with four or more stimuli if they do not take subjects long to process, such as a short story or photograph. For stimuli that take longer to process, start with three. The advice to "start with" a certain number is because sometimes the different stimuli do not always adequately represent the concept or level of the independent variable one is attempting to create. That is the subject of the next section of this chapter. Of course, it is never possible to make multiple stimuli exactly the same on every possible category, so message variance is treated as random error when testing the differences between levels of treatments.[56]

Even though creating more stimuli is more work, it pays off in the long run, as experiments that use single-message designs will eventually need to be conducted again and again with different stimuli and then meta-analyzed in order to be generalized.[57]

For Review-No Commercial Use(2023)

Fixed vs. Random Factors

When multiple stimuli are used to represent a single level of a factor, the standard way to treat the various responses is by averaging each individual subject's responses into all the stimuli that represent each level.[58] So, for example, if a subject sees three photographs of Black people and three photographs of White people and is asked to make an ethical decision, then the subjects' ethics scores on three Black photographs are averaged, resulting in one score for that level of the race factor, and the same done for the three White photographs. These are then analyzed with traditional statistical techniques such as analysis of variance (ANOVA).[59] This is appropriate when the levels of the factors are considered **fixed factors** rather than **random factors**. Fixed factors are those whose levels are limited,[60] or researchers choose the levels,[61] whereas random factors are those that represent a sample of a larger population rather than every conceivable level within that population.[62] The most common use of a random factor is with subjects of an experiment who are samples from a larger population and do not exhaust the universe of potential subjects.[63]

Sometimes, the stimuli also should be treated as random factors. Using the example of a study whose stimuli are photographs of Black people and White people, the individuals in

the photographs are obviously not the only Black and White people in the world but are sampled from a larger universe. If researchers have sampled these particular photographs and use them to generalize to a larger population of all people of those two races, those photographs would be considered random factors. In such a study, both the subjects who participate in the study and the stimulus photographs are random factors, and special statistical techniques for mixed-effects models should be used.[64] This allows for generalizing to the population of stimuli as well as to the population of subjects.[65]

Using random effects models also helps avoid making a Type 1 error, which was described in chapter 8 as a "false positive," or when one finds an effect that is not really real.[66] These special statistical techniques are not commonly used in some social science fields, or even in clinical and social psychology.[67] Random effects models are recommended when there is an infinite number of levels or many more than are being tested and the levels of the factor are sampled. This allows for generalizing beyond the particular levels in the experiment.[68]

In contrast, a fixed factor has a limited number of levels—a drug that exists only in certain dosages, such as 10 mg, 20 mg, and 40 mg, for instance. A drug with an infinite number of levels would be a random factor. If only two interventions exist—for example, there are only two training programs for social workers dealing with violence-prone clients—that is a fixed factor. A study comparing these different tests when other tests exist is another example of a fixed factor. Another common use of fixed factors is with panel data that compare the same people, companies, states, schools, or other entities **longitudinally**, or across some longer period of time. A fixed factor also can be the levels of a factor that are chosen by the researcher. For example, it is common to treat stimuli such as the photographs of Black and White people as fixed factors when the researcher is only interested in generalizing to the specific stimuli used in the experiment, not beyond. This can lead to the necessity to replicate findings with different stimuli so as to aid generalizability.

For Review-No Commercial Use(2023)

The discussion surrounding the appropriate use of fixed vs. random effects models is complicated and beyond the scope of this book. Interested readers will find several articles for further information in the "Suggested Readings" section.

MANIPULATION CHECKS

Once the stimuli are created, it is often advisable to conduct a manipulation check to determine if subjects perceive the stimuli the way researchers intend.[69] Failure to ensure that a stimulus is what is intended in concrete terms can doom an experiment to failure, specifically of the Type 2 variety, which is not supporting a hypothesis when it should. Scientific knowledge is not advanced and doubt may be cast upon theories that are

actually correct when this happens. In addition, researchers are less likely to publish studies that result in null findings.[70] It has been suggested that manipulation checks be added to the checklist of requirements for experiments in numerous social sciences,[71] with some such as social psychology requiring them before a paper is accepted for publication.[72] However, this is a contested point, with some saying that manipulation checks are rarely, if ever, necessary.[73]

Manipulation Check, Pretest, or Pilot Study?

A manipulation check is a way of determining that the stimuli have actually manipulated the independent variable the way the researcher intends.[74] The terms *pilot study* and *pretest* are frequently used interchangeably with *manipulation check*. Often when researchers conduct the manipulation before the actual study, they will refer to it as a pretest. This book delineates the three as distinct procedures. In this text, a pretest, as in "pretest–posttest design," is part of the actual study, given before the treatment or manipulation, and is either equivalent to or the same as the posttest so before-and-after comparisons can be made. A pretest is different from a manipulation check because their purposes are different, with a manipulation check designed to determine if the researcher's manipulation of latent constructs have been done effectively, no matter when it is carried out. Chapter 11 discusses pilot studies in detail, but briefly, a pilot study is a larger undertaking than a manipulation check, intended to be a "test run" of the entire experiment.

For Review-No Commercial Use(2023)

Manipulation checks are conducted to help ensure that stimuli are what the researcher intends them to be so as to aid the best possible chance of having an effect on the theoretically relevant outcome variables.[75] For example, in a study to determine if vivid writing had an effect on moral judgment,[76] the manipulation check was designed to see if the three stories the researchers had written were "vivid," which was defined theoretically as concrete, descriptive writing that helped people paint a mental picture. The manipulation check determined if ordinary people found the vivid stories to be more concrete, descriptive, and mental-picture-painting than the control stories written in traditional news style. This was achieved by asking subjects who were not involved in the actual experiment to answer questions specifically about the vividness of the stories rather than to take the entire study. Some of the questions to that effect were: "This story was very descriptive," "This story was very colorful," "This story was very vivid," "This story had a lot of details," and others designed to measure the features that the literature said defined the theoretical construct of vividness. In this manipulation check, one of the four stories was not significantly different from the nonvivid stories, so it was rewritten and checked again before being used with subjects for the actual experiment. In this case, the manipulation check

uncovered a problem so that the researchers could fix it before it was too late. A manipulation check is not a measure of what the stimulus does in terms of effects on the dependent variable; rather, it is a measure of what the stimulus is.[ii]

Direct vs. Indirect Manipulations

In the study on vivid writing, the construct of vividness could only be manipulated indirectly—for example, with descriptions of people and scenes. A manipulation check is needed for this kind of latent independent variable in order to know if the researchers have successfully created vivid stories.[77] This is not the case for all manipulated variables; for example, in a study where the manipulation is the number of sources in a story, a manipulation check is not necessary if one group of stories contains five sources and the other contains one source. In this case, the operationalization is the same as the construct—there is no need to conduct a manipulation check to see if some stories have five sources and some have one; it is indisputable that they do. "So long as the operationalization of the treatment and the independent variable are one and the same, there was no need for a manipulation check."[78] Similarly, gender may not need to be subjected to a manipulation check if names and physical features are not ambiguous. For example, in one study, researchers used the names "John" and "Sarah" in short profiles of hypothetical political candidates.[79] Nor would political party affiliation need a manipulation check as long as it was clearly labeled. Concrete attributes of the stimuli or intrinsic structural properties do not always need a manipulation check. For example, it is not necessary to ask subjects how long two stories are if one is 500 words and one is 100 words. No matter what subjects say, they clearly are different lengths.[80]

For Review-No Commercial Use(2023)

Mediators and Psychological States

Manipulation checks should be conducted not on the indisputable structural features of the stimuli but on the psychological states that these features arouse. For example, in the vividness study, in addition to ensuring that the stories were colorful, descriptive, and contained details—all content features—the manipulation check also ensured that the stories were constructed to have the desired psychological effect on subjects—that is, that they helped them paint a mental picture, a theoretical construct called "imaging."[81] The manipulation check also included questions about this, including, "I formed a mental image from this story," "It was easy to imagine what this scene looked like," and "I could easily visualize this."

[ii]There is some dispute over what researchers call manipulation checks actually being measures of mediating psychological states. Manipulation checks measure structural or content attributes of the stimuli rather than attitudes or perceptions that subjects have of the content; these are attributes of the participant rather than attributes of the stimulus. A detailed discussion of this is beyond the scope of this book, but interested readers can learn more in O'Keefe, "Message Properties, Mediating States, and Manipulation Checks."

These represent the subjects' perception of the stories as vivid or not, which is theorized to induce the psychological state of imaging, or painting a mental picture. The psychological state represents the causal explanation referred to in chapter 1. In many cases, researchers use psychological states in manipulation checks but then fail to analyze them in the actual study as potential mediators of the dependent variable.[82] When this is the case, researchers miss an opportunity to provide an explanation for an effect. Including potential mediators in studies enables researchers to explain why a certain effect occurred (see More About box 9.3, Mediators and Moderators). For example, in one study, researchers examined the effect of the number of terms or phrases used in a web search on users' perceptions of control or mastery, a concept called *dominance*.[83] Dominance was the dependent variable (DV) and Search Term was the independent variable (IV) represented by two levels, high and low, operationalized by three, and one search term or phrase. The search terms or phrases are physical attributes of the stimuli. The psychological state, or mediator, that the number of search terms were expected to activate in subjects was perceived relevance; three search terms were intended to be perceived by subjects as resulting in highly relevant results, and one search term was expected to return a search result that subjects perceived as low in relevance. The researchers conducted a manipulation check to ensure that three search terms led to subjects' perceiving results to be highly relevant and one search term leading to results perceived to be of low relevance. This study also included a measure of measuring perceived relevance in the actual study so that it could be analyzed as a mediator; that is, it was hypothesized that perceived relevance was the causal mechanism that would mediate the relationship between the number of search terms and dominance.[iii] In this case, three hypotheses are proposed: one proposing the direct effect of the IV (number of search terms) on the DV; one proposing the direct effect of the mediator (perceived relevance) on the DV; and one proposing the effect of the IV, indirectly through the mediator, on the DV.

For Review-No Commercial Use (2023)

One argument for conducting a manipulation check is that without it, the scientific community loses valuable information. When no manipulation check is done in a study that fails to support a hypothesis, researchers cannot know if the theory needs revision or if it was the treatment that was unsuccessful. The hypothesis may even be abandoned.[93] As Mutz and Pemantle say, "Ignoring manipulation checks thus impedes the growth of scientific knowledge."[94]

[iii]It is beyond the scope of this book to discuss the analysis of mediators beyond saying they can be tested using regression as described by Baron and Kenny (1986), or with bootstrap distribution tests as described by Efron and Tibshirani (1993) via a free macro for certain statistical software called PROCESS (Hayes, 2013). It is available at http://www.processmacro.org/index.html. R. M. Baron and D. A. Kenny, "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* 51, no. 6 (December 1986): 1173–1182; Andrew F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach* (New York: Guilford Press, 2013).

MORE ABOUT . . . BOX 9.3
Mediators and Moderators

Chapter 1 discussed the importance of including a causal mechanism in an experiment, which is something to explain *why* the effect an experiment found actually occurred. The example given there was of an experiment to determine if photographs improved moral judgment. In that study, photographs did improve moral judgment, and that was the effect. The causal mechanisms examined were involvement and elaboration, or being absorbed or interested in the stories and thinking deeply about the people in them. These causal mechanisms were found to be mediators—that is, they were the psychological states that served as the pathways by which the independent variable influenced the dependent variable.[84] There is often confusion about when something is a mediator and when it is a moderator. This supplement is designed to help sort that out.

Mediators

Mediators are psychological states or conditions that explain how and why the observed effect happened.[85] They represent the mechanisms by which some effect occurs.[86] In James Lind's experiment on scurvy described in chapter 2, he found that citrus like limes and lemons cured scurvy. He did not study the mediators or mechanisms in that study, but later it was found that vitamin C was the mediating variable between limes and scurvy. In the earlier example, central route processing or thinking deeply about the photographs was the cognitive process or psychological condition that led to better moral judgment. Fear could be the psychological state that leads to intentions to quit smoking after seeing PSAs. In addition to representing psychological states, mediators come after the treatment or manipulation is given. Continuing the previous examples, fear came after watching the PSA, while elaboration was induced by seeing the photograph. Finally, the mediator must be correlated with the independent variable. Thus, the PSAs themselves had scary messages, which correlate with subjects' feelings of fear, and the photographs had elements of novelty and unusualness, which has been correlated with greater cognitive elaboration. Treatments probably have many mediators or causal mechanisms, but the only ones that researchers can know about are the ones they test.

For Review-No Commercial Use(2023)

Moderators

Unlike mediators, moderators explain when and under what conditions effects might be larger or smaller, or when and for whom certain effects will be found.[87] A moderator is a third variable that, when combined with the independent variable, has an effect on the strength of the relationship to the dependent variable. Examples of moderators include older age decreasing the likelihood of finding a job; attending a preschool moderating the effect of a mother's age on a child's cognitive development;[88] and personality traits such as hypermasculinity increasing aggression after viewing violent

TV shows.[89] In all of these examples, the moderators are observable traits that are preexisting, or come before the treatment, rather than being induced by the treatment. For example, age and a hypermasculine personality are traits or characteristics of the job applicant and TV viewer, not something induced by a treatment such as a job training program or watching violent TV. Also, moderators are uncorrelated with the independent variable or manipulation. For example, whether a child attends preschool is not correlated with the mother's age.

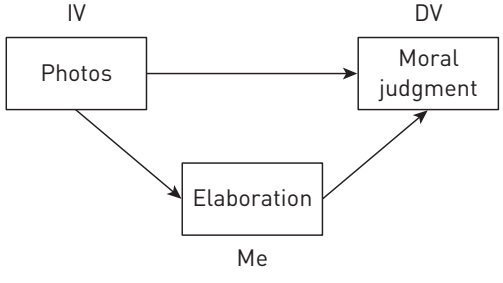
The following chart summarizes the differences between mediators and moderators.

Mediator	Moderator
Explains why and how an effect occurred; the causal mechanism	Answers for whom and when the effect occurs; affects the strength of the relationship between the IV and DV
Is a psychological state	Is an observable trait
Comes after the IV	Comes before the IV
Is correlated with the IV	Is uncorrelated with the IV

For Review-No Commercial Use(2023)

In addition, there is a graphic display that helps visualize the relationship between mediators and moderators in an experiment. This graphic is frequently found in journal articles to represent the processes of an experiment. Mediation is identified with the abbreviation Me, moderation with the abbreviation Mo. The following example uses photographs as the IV and moral judgment as the DV. In the mediator graphic, elaboration, or thinking deeply, is the mediator, while education level is the moderator. The arrows represent the process and also the different hypotheses, printed underneath.

Mediation graphic



(Continued)

(Continued)

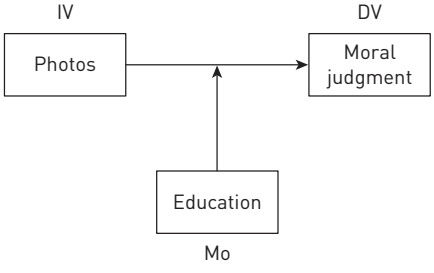
H1: Seeing photographs will significantly increase the level of moral judgment compared to not seeing photographs.

H2: Seeing photographs will lead to greater elaboration.

H3: Elaboration will lead to significantly higher levels of moral judgment.

H4: Seeing photographs will lead to higher levels of moral judgment, mediated by elaboration.

Moderation graphic



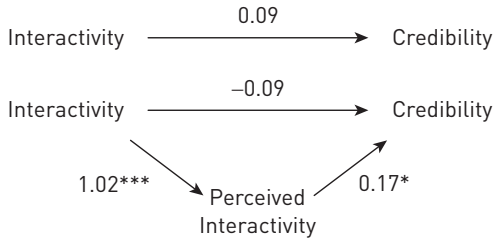
For Review-No Commercial Use(2023)

H1: Photographs will significantly increase levels of moral judgment.

H2: Education moderates the effect of photographs on moral judgment, with higher education leading to higher levels of moral judgment.

Here is an example from a published study of the graphic display for a mediator and the hypotheses.[90] This experiment tested the effect of interactive features of a website on subjects' assessments of the credibility of the news on the website. The IV was interactivity, with the treatment group instructed to use the interactive features, operationalized as voting in an online poll, e-mailing the news organization, and viewing a slide show. The noninteractive group subjects were asked to read three articles. The mediator was the psychological state of perceived interactivity—that is, the causal mechanism was how interactive subjects perceived the website to be. You can see the elements that define a mediator here: Perceived interactivity comes after the treatment is given; that is, being presented with online polls, e-mail, and a slide show should lead subjects to perceive the website as interactive. It is correlated with the IV, in that interactive features are related to perceptions of interactivity. Perceived interactivity is also a psychological state—that is, it is a perception. And it explains why the IV has an effect on the DV; giving people websites with polls, e-mail, and slide shows leads them to perceive it as interactive, which increases their assessment of the site as credible.

From: Tao, C., and E. P. Bucy. 2007. "Conceptualizing Media Stimuli in Experimental Research: Psychological Versus Attributed-Based Definitions." *Human Communication Research* 33: 397–426.



H1: Credibility ratings of online news sites will be higher for interactive conditions than noninteractive conditions.

H2: Perceived interactivity will be positively associated with assessments of media credibility.

H3: Perceived interactivity will mediate the relationship between interactive tasks and evaluations of media credibility.[91]

For Review-No Commercial Use(2023)

It is beyond the scope of this textbook to explain how to analyze mediators and moderators. The classic way uses a series of regressions as explained in step-by-step detail by Baron and Kenny, and the Sobel Test; newer procedures use bootstrapping and can be accomplished using a free macro called PROCESS.[92] It is available at http://www.processmacro.org/index.html.

Manipulation Checks in Survey Experiments

The trend of using survey experiments also argues for the need for manipulation checks.[95] With a survey experiment where subjects are taking the experiment at remote locations where the researcher cannot observe, it is impossible to know if subjects were actually exposed to the stimuli. Embedding a manipulation check into the actual experiment can help researchers identify subjects who paid attention to the stimuli versus those who

randomly checked off answers.[96] Screener questions are not a substitute for manipulation checks in survey experiments because screeners are designed to test if subjects were merely exposed to the manipulation, not whether it was effective.[97]

One last reason to conduct a manipulation check is to determine if the stimuli are similar on things that are not intended to vary and could confound the results. For example, in designing multiple advertisements, experiments might use different models who should be similar in appearance and attractiveness; stories should be similarly interesting or likeable; political ads should be similarly toned, whether negative or positive. Manipulation checks can confirm these similarities so that confounds are not the cause of results.

Many experiments do not report manipulation checks,[98] leading readers to believe that none were conducted. Some articles say that a manipulation check was conducted and that it worked but provide no supporting evidence. It is becoming more common to see details about manipulation checks provided in appendices or in online supplements. Manipulation checks may not need to be conducted when the manipulations or stimuli have been used previously, in which case that is mentioned.

For Review-No Commercial Use(2023)

Practical Issues

Three practical concerns about manipulation checks are when to conduct them, how many subjects to use, and how to measure things.

There are two options for when to conduct a manipulation check: before the actual study and separate from it, or embedded within the actual study with its assessment placed at the end. Actually, there are three options, as manipulation checks can be done both ahead of the study and also within it. This can be the preferred option, which will be explained next.

Conducting a manipulation check before running the actual study is a helpful way to design successful stimuli. It allows researchers to spot problems and tweak the stimuli before testing it on actual subjects. Two researchers did this in their political study, using students for the manipulation check to validate the stimuli but not administering it to the real politicians in the actual study so as not to make it too lengthy.[99] For this type of a manipulation check during stimulus development, fewer subjects are needed than for the actual study, but still enough to find significant differences between the treatment and control group stimuli. The study on vivid writing used twenty-six subjects for the manipulation check conducted prior to the actual study, for example.[100] But this is probably the least that should be used, as fewer than twenty subjects can make it hard to conduct significance tests.

Another common way that social science researchers report manipulation checks is embedded within the study itself. The advantage of this is that they are assured the actual subjects of the experiment perceived the stimuli manipulations as they were intended. This should be done with the manipulation check assessment asked at the end of the study, after the dependent variable measures are given, so as not to alert subjects to the manipulations or purpose of the study.[101] This also has the benefit of measuring the psychological state expected to cause the outcome in the actual study, affording researchers the ability to test mediators.

The third way to conduct manipulation checks combines the two approaches: a separate group of subjects who are not involved in the study and then another manipulation check within the study itself so that actual experimental subjects also receive it. The benefits to this are threefold: this allows researchers time to revise the stimuli if necessary, ensures that actual experimental subjects perceive the stimuli as intended, and incorporates a mediator into the study. The only cost to doing this is that it lengthens the study by the amount of time subjects spend on the manipulation check questions.[102] One way around this is to reduce the number of manipulation check questions for the actual study based on an internal consistency analysis such as Cronbach's alpha or factor analysis, and then eliminate the weaker items from the actual study. The vivid writing study described earlier used nine items assessing vividness in the first manipulation check and reduced the number of questions to the three most predictive items for the actual study. Many researchers' experiences with stimuli that fail manipulation checks and need to be revised, and even prestudy manipulation checks, go on behind the scenes and do not get reported in journal articles because of word-length limits. Readers should not despair; research can be messy, and what appears in print is not necessarily the whole story.

For Review-No Commercial Use(2023)

When measuring for a manipulation check, one does not ask all the same questions as in the study itself; the goal of the manipulation check is to determine if the operationalization reflects theoretical constructs, so assessments should focus on that, not on measuring outcome DVs. As always, common sense and the literature should guide the writing of manipulation check questions. If other researchers have already conducted manipulation checks on the construct in a study, that is a natural place to begin before attempting to create one's own assessment. But when none can be found, it is perfectly acceptable to create one using the literature as a basis.

If a manipulation check embedded within the actual study fails—that is, if subjects do not perceive it the way it was intended—that stimulus should be dropped from the analysis if it cannot be revised and rerun.

REPORTING THE STIMULI AND MANIPULATION CHECKS

Stimuli Write-Up

When writing the paper for an experiment, a portion of the methods section, usually labeled *stimuli*, *stimulus materials*, or *procedures*, is where information about the stimuli and any issues associated with them are presented. This should appear early in the methods section and describe the stimuli in enough detail for readers to visualize them. It also addresses any issues that might be of concern, such as the fictitious nature of people, organizations, etc.

For experiments where the details of the stimuli are not too extensive, they should be reported in the body of the paper, including the exact wording of all the messages and description of any visual images. In the marketing experiment on cookstoves, the posters' wording was short, so it was all contained in the text, one of which is reproduced here:

> Smoke from the cook fire is poison. It makes you feel light-headed or dizzy, makes you cough, and can cause sore eyes or a sore throat from the smoke. Smoke from cook-stoves causes serious diseases, including pneumonia and bronchitis. These diseases from cook-stove smoke caused as many child deaths in Uganda as malaria.[103]

For Review-No Commercial Use(2023)

The text also describes the picture on the poster as a baby smoking a cigarette.

For studies with many stimuli or stimuli that are more complicated to describe, it is appropriate to give the full description of one stimulus as an example and refer readers to an appendix or a website for the others.[104] It is also a good idea to make it easy for readers to see what the manipulations were by highlighting or italicizing the items that vary while using regular type for the parts that are constant. For example, in the study of parliamentarians who made it into the news, one of the twenty-four press releases was described this way:

> "The B61-nuclear bombs do not need to be modernized, but rather destroyed," *responds Green Member* of Parliament *Wouter Devriendt* to the recent news about *the modernization of the nuclear weapons stored in Kleine Brogel.* "Nuclear weapons are dangerous and useless. Moreover, the modernization, the storage, the maintenance and the surveillance of the nuclear bombs are extremely expensive. The government needs to undertake action to commence and finish a complete nuclear disarmament." Devriendt wants to gain clarity about the measures concerning nuclear weapons in Belgium *by asking a written question* to the authorized cabinet member.[105]

Finally, for very complex stimuli and when there are many of them, the text should describe the stimuli briefly and then refer readers to an appendix or supplemental website where they can read and view the entire body of stimuli. In the study of corporate social responsibility (CSR), for example, the stimuli section gave only a general description of the scenarios and the dimensions that they represented, such as whether the corporation was harmful to the environment or not, a good business partner, and how it treated employees.[106] Because there were six lengthy text scenarios, presenting it in the appendix was more efficient than in the text. The description in the body of the paper was detailed enough that most readers were not left with many questions. Here is the text:

> For this task, six different scenarios were created and distributed to the participants. In each scenario, the participants were provided with information on a fictive organization operating university canteens. In accordance with Bardsley (2000), the participants were not informed that the survey subject was fictional as this does not constitute a deception of participants, and the scenarios could potentially be true. In scenarios 1, 3, and 5, the organization was described as a "for-profit organization" and in scenarios 2, 4, and 6 as a "nonprofit organization." The scenarios also differed in terms of the organization's CSR performance. In the first two scenarios, no information was given about CSR performance. Scenarios 3 and 4 contained information about the organization's positive CSR performance and scenarios 5 and 6 provided information about negative CSR performance. The CSR performance was manipulated by providing information on organizational behavior along three fundamental dimensions of CSR: The environment (environmentally friendly vs. environmentally harmful waste disposal), society (respectful vs. disrespectful treatment of employees), and economy (pleasant vs. unpleasant business partner). . . . The original texts describing each scenario can be found in the Appendix.[107]

For Review-No Commercial Use(2023)

Elements to be included in the written description of the stimuli include, in no particular order,

- the number of stimuli created and the number in each condition;
- if the stimuli were real or fictitious:
  - if fictitious, describe how subjects were persuaded they were real, and
  - if real, describe how subjects' prior knowledge was not a confounding factor;
- if deception was involved, how subjects were debriefed;
- how the concepts of theoretical interest were operationalized;

- the physical attributes of the stimuli that are posited to cause the changes in the dependent variable (i.e., pictures, format, length, word count, etc.);
- discussion of any potential confounds, whether of the stimuli or procedure, linked to the independent variables and how they were controlled (e.g., statistically or in the manipulation);
- how differences were maximized; and
- explanation of all decisions in theoretical terms—for example, if gender has been found to be related to a particular outcome or not, and how the operationalization reflects theoretical constructs.

For more examples of writing the stimuli portion of the methods section, see How To Do It box 9.4.

One piece of advice for writing up the stimuli section is to do it immediately after creating the stimuli. It may be hard to remember everything that was done or the reasoning behind the decisions that were made after the study is completed. The same advice goes for writing up the manipulation check results; rerunning manipulation check data because the original output could not be found is unnecessarily time consuming. Develop a habit of writing each portion of the methods as soon as it is finished, even if it is just in note format.

For Review-No Commercial Use(2023)

Manipulation Check Write-Up

Some journals put the manipulation check in the methods section, immediately after the stimuli are described, while others save it for the results section, especially if the manipulation check was conducted on the subjects of the actual study. Some studies do a combination of both, reporting the measures used in the manipulation check in the methods section and then the test of those measures reported in the results section[108] (see How To Do It box 9.5, Writing Up Manipulation Check Results). If the journal has no preference, then it should be included where it makes the most logical sense and will not leave readers with lingering questions. Another option for very long manipulation checks, or when checks were done before and during the actual experiment, is to include the results in an appendix or refer readers to a supplemental website.

HOW TO DO IT 9.4

Writing Up the Stimuli

Here are some examples of how researchers have described their stimuli. The description should be detailed enough for readers to visualize what the stimuli looked like and to be able to recreate them for their own study.

Example 1

In this study, the description of the stimuli is combined with the details of the study design, covered in chapter 6. Notice also how this addresses the use of fictitious stimuli and debriefing of subjects regarding deception, and also how it discusses confounds and how they were addressed:

"Methodology

Design and stimuli

Study 1 used a single-factor design in order to examine the impact of lowered and raised voice pitch on the perceptions of an organizational spokesperson. Each participant first received a small text online informing them of the crisis events that were taking place in a fictitious company, followed by the voice recording of an organizational spokesperson. The use of a fictitious organization allows us to avoid potential confounding effects of pre-existing knowledge and prior perpetration (Laufer & Jung, 2010). However, to avoid suspicion among participants due to the fact that they had not heard of the organization or the events, the crisis was situated in a different region of the country. At the end of the experiment, the respondents were notified that the crisis was fictitious.

For Review-No Commercial Use(2023)

The introductory text explained that a company, which cultivates salad, was at that moment confronted with a food contamination crisis. Consumers who had eaten the salad were reporting to hospitals with stomach problems. In addition, it appeared that the salad got contaminated because the company had used a pesticide which was forbidden by the local government. The crisis thus posed a threat to the public, which increased the demand for an adequate organizational response.

Voice pitch was manipulated by means of a brief sound fragment from a fictitious press conference given by the organization in crisis. An organizational spokesperson confirmed the information that had been offered in the introductory text, informed potential victims, elaborated briefly on the corrective action that was taken, and took responsibility and apologized for the crisis. We audio-recorded a male student in radio journalism while reading this statement about the crisis events. Prior research found no impact of gender on the relation between voice pitch and attributions of competence and dominance (DeGroot & Motowidlo, 1999; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010; McHenry et al., 2012). Similar to previous studies, we manipulated the voice pitch in the recordings by means of computer software (i.e., Audacity) (Imhof, 2010; Jones et al., 2010; O'Connor et al.,

(Continued)

(Continued)

2011; Puts et al., 2006, 2007). The software allowed us to manipulate one specific aspect of the voice, namely pitch, while keeping all others (e.g., speed, emphasis, pausing, accent, and pronunciation) constant. This method also allowed us to control for content (Imhof, 2010).

Two versions of the original voice recordings were made, one with raised voice pitch and one with lowered voice pitch (cf. Jones et al., 2010; Puts et al., 2006; Tigue et al., 2012). The baseline level was raised and lowered by one-and-a-half semitone (cf. Puts et al., 2006, 2007)."

From: Claeys, An-Sofie, and Verolien Cauberghe. 2014. "Keeping Control: The Importance of Nonverbal Expressions of Power by Organizational Spokespersons in Time of Crisis." Journal of Communication **64: 1162–1163.**

Example 2

The next two examples are particularly concerned with making the stimuli as realistic as possible. In the first write-up of stimuli, notice how the researchers report adapting actual messages for the study:

"Stimulus materials

An array of advertorials and advertisements were collected from media sources to generate the stimulus materials for Study 2. Each advertorial contained both a text-based information section focusing on healthy eating or sleeping and an advertising section presenting visuals with advertising captions. The delivered information on healthy eating and sleeping was similar to the information used in Study 1. Using visual editing programs, we separated and placed informational content before advertising content for info-first advertorial stimuli, and placed advertising sections ahead of informational sections for ad-first advertorial stimuli. Except for this experimental manipulation in structure, the messages and conveyed information were identical. Advertising-only sections were used as regular advertisements (see Appendix A for sample stimulus materials)."

For Review-No Commercial Use(2023)

From: Kim, Sunny Jung, and Jeffrey T. Hancock. 2016. "How Advertorials Deactivate Advertising Schema: MTurk-Based Experiments to Examine Persuasion Tactics and Outcomes in Health Advertisements." Communication Research **44 (7): 1019–1045. 10.1177/0093650216644017**.

Example 3

In this example, the researchers used an expert to help them create realistic stimuli. Notice also the informal "manipulation check" in this study from 2003; today, "squaring nicely" is unlikely to pass the test for adequate reporting of manipulation check results:

"First, we ran copies of our ads by a local political consultant to see if they were realistic. He indicated that the spots were plausible representations of the kind of ads one might hear run during an election, which gives us greater confidence in the generalizability of our results. Second, we asked students from one of the author's university to rate how negative and how positive were each of the simulated ads. We wanted to make sure that our ads captured accurately the variables of interest. The undergraduates' ratings of the ads squared nicely with our judgments of them. In other words, students viewed our 'attack' ads as attacks and viewed our 'positive' ads as positive. This verification increased our confidence in the manipulation."

From: Geer, John G., and James H. Geer. 2003. "Remembering Attack Ads: An Experimental Investigation of Radio." Political Behavior **25 (1): 69–95.**

HOW TO DO IT 9.5

Writing Up Manipulation Check Results

Manipulation checks are reported in different sections of a study, depending on the journal's preference and where it makes the most sense. Here are some examples.

Example 1

In this study, the manipulation check results were reported in the methods section after the stimulus material, and said:

"On a seven-point scale (1 = strongly disagree to 7 = strongly agree), participants were asked to what extent the article dealt with the (1) positive or the (2) negative consequences of the issue. The manipulation check showed successful manipulation with slightly elevated standard deviations. When asked for positive consequences, participants in the opportunity condition ($M = 5.94$, $SD = 1.63$) perceived their article to be more positive than participants in the risk condition ($M = 2.35$, $SD = 1.93$; $t(612) = 2.75$, $p < .001$). When asked for negative consequences, participants in the risk condition ($M = 5.87$, $SD = 0.146$) perceived the article to be more negative than participants in the opportunity condition ($M = 3.24$, $SD = 2.09$; $t(611) = 4.53$, $p < .001$)."

From: Lecheler, Sophie, and Claes H. de Vreese. 2013. "What a Difference a Day Makes? The Effects of Repetitive and Competitive News Framing Over Time." ***Communication Research*** **40 (2): 157.**

For Review-No Commercial Use(2023)

Example 2

In this study, the manipulation check was reported in the results section:

"Using condition (action exemplar, inaction exemplar, and control) as the independent variable and negative affect as dependent variable, a one-way analysis of variance (ANOVA) reported a significant effect, $F(2, 192) = 9.24$, $p < .001$, $\eta 2 = .09$. A Bonferroni post hoc test reported that those in the action exemplar condition, $M = 3.56$, $SD = 1.2$, reported higher negative affect scores than those in the control condition, $M = 2.85$, $SD = 1.33$, $p < .01$, and inaction exemplar condition, $M = 2.62$, $SD = 1.2$, $p < .001$. Therefore, the manipulation for the action risk exemplar worked.

However, those in the inaction exemplar condition produced negative affect scores no different than the control condition, $p = .95$, indicating that the manipulation for the inaction exemplar condition did not achieve the desired result."

From: Dixon, Graham N. 2015. "Negative Affect as a Mechanism of Exemplification Effects." ***Communication Research*** **43 (6): 768**

Example 3

In another study, the manipulation check results were included in a supplementary online site; the text gave this summary of the results:

"The priming manipulations were pretested for their effect on various associations and discrete emotions. Results confirmed that the manipulations are effective. Thus, t-tests indicate that the harm prime generated significantly more injury and harm associations compared to the other primes, that

(Continued)

(Continued)

the damage prime generated associations with convenience, and the disgust and sadness primes each generated the relevant emotion significantly more than the other three primes. Further, primes were overall comparable in the level of perceived negativity (the only pair of essays that significantly differed on a t-test were disgust and damage; see the online supporting materials for full pretest results)."

From: Ben-Nun Bloom, Pazit. 2014. "Disgust, Harm and Morality in Politics." ***Political Psychology*** **35 (4): 501.**

Example 4

Finally, in this study, part of the manipulation check was reported in the methods section and part in the results:

In Methods

"Measures were rated on 7-point Likert scales except where noted as otherwise.

Manipulation check.

At both pretest and posttest perceived threat was assessed using items adapted from prior studies (Rimal et al., 2009; Tyler & Cook, 1984; Witte, 1994). To measure perceived susceptibility, participants were asked how likely it was that they themselves or others around them might be exposed to HIV/AIDS. Participants were also asked how serious and threatening HIV/AIDS was and how important or urgent HIV/AIDS prevention was to individuals and to society. Responses were averaged into indices of perceived severity ($\alpha = .86$ at pretest, $\alpha = .95$ at posttest).

For Review-No Commercial Use(2023)

And in Results

"From pretest to posttest, participants with main partners perceived higher susceptibility at the personal level, $M = .50$, $SD = 1.34$, $t(157) = 4.68$, $p < .001$; and at the societal level, $M = .37$, $SD = 1.23$, $t(157) = 3.88$, $p < .001$. From pretest to posttest, participants with nonmain partners perceived higher susceptibility at the personal level, $M = .47$, $SD = 1.38$), $t(111) = 3.62$, $p < .001$; and at the societal level, $M = .43$, $SD = 1.30$, $t(111) = 3.49$, $p < .001$. Perceived severity did not increase from pretest to posttest. Hypothesis 1 was, therefore, partially supported. The manipulated difference in perceived threat was reflected by perceived susceptibility."

From: Zhang, Jueman (Mandy), Di Zhang, and T. Makana Chock. 2014. "Effects of HIV/AIDS Public Service Announcements on Attitude and Behavior: Interplay of Perceived Threat and Self-Efficacy." ***Social Behavior and Personality*** **42 (5): 802–803.**

Regardless of where it appears, elements that should be included in the manipulation check description include:

- if manipulation checks were conducted prior to the actual study, within the study itself, or both;
- a statement of whether the manipulation was successful or not;

- *N*s, statistical tests, and values that indicate significant differences and the direction of the differences;
- the measures used and response sets; and
- if items were created by others (cite them) or the researchers.

With stimuli created and deemed successful via a manipulation check, the design of the experiment is nearing completion. The next chapter deals with construction of the instrument used to measure the dependent variables, covariates, mediators, and moderators. Of course, the steps in experiment design can be reversed, with instrument design coming before stimuli construction, or be done at the same time.

Common Mistakes

- Not using stimuli that look realistic
- Using one stimulus rather than many
- Not doing a manipulation check when necessary, or not reporting it
- Underestimating the strength of a manipulation needed to produce results
- Not including psychological states as mediators in the actual study

For Review-No Commercial Use(2023)

Test Your Knowledge

1. In a study, journalists read press releases that highlighted different features of politicians, such as their political party and favorite issues, and were asked how likely they would be to write a story about them. What was the category of theoretical interest manipulated in the stimuli?
   a. The likelihood that journalists would write stories
   b. Different features of politicians that make them newsworthy
   c. The length of the press releases
   d. The size of the news organization the journalist worked for
2. Stimuli represent categories of theoretical interest because they _____.
   a. Serve as the vehicle through which the experimenters manipulate the factors hypothesized to cause some effect or outcome
   b. Are complex and take a lot of work to create

c. Control for confounding variables
d. Measure and statistically control for confounds

3. In an experiment that uses real politicians in the stimulus materials, it is important to control for subjects' prior knowledge about the politicians.
   a. True
   b. False

4. A statistical control is _____.
   a. Something that cannot be controlled in the stimuli
   b. Something that is measured and controlled for
   c. Both A and B
   d. A statistical technique like ANOVA or chi-square

5. Statistical controls should be measured _____.
   a. Before the stimuli are given
   b. After the stimuli are given
   c. During the administration of the stimuli
   d. Two weeks after the study

For Review-No Commercial Use(2023)

6. Fictitious people, places, and events are often used in experiments to avoid contamination by subjects' prior knowledge. These are called _____.
   a. Statistical controls
   b. Experimental controls
   c. Control groups
   d. Treatments

7. Selecting political issues that are not clearly linked with a particular party represents a form of _____.
   a. Statistical control
   b. Experimental control
   c. Control group
   d. Treatment

8. Covariates should _____.
   a. Always include demographics
   b. Be included in the manipulation check
   c. Be randomly assigned
   d. Be related to the outcome or dependent variable

9. Fixed factors are _____.
   a. Those whose levels are unlimited
   b. Those whose levels are chosen by the researchers
   c. Those that represent a sample of the population
   d. Both A and B.
10. Random factors are _____.
   a. Those whose levels are unlimited
   b. Those whose levels are chosen by the researchers
   c. Those that represent a sample of the population
   d. Those that are randomly assigned

Answers

1. b
2. a
3. a
4. c
5. a
6. b
7. b
8. d
9. d
10. c

For Review-No Commercial Use(2023)

Application Exercises

1. For the experiment you are designing in this course, describe how you will create your stimuli. Address the following:

   - **How will you make it realistic?** For example, will you start with an actual website or stimuli and adapt it? Do you have professional experience designing ads? Will you employ someone who does? Will you use software or online tools? Or will you use existing stimuli such as a TV show or curriculum plan?
   - **How do your stimuli represent the categories of theoretical interest?** For example, do comments at the end of an online story represent the theoretical category of reader engagement?
   - **How will your stimuli maximize comparisons?** Are you employing the strongest manipulation or the largest difference possible?
   - **How will you employ message variance?** That is, how many stimuli will you create? Why?

   Once you have planned all the details of your stimuli, begin creating them.

2. Design a manipulation check for the stimuli you have created. What questions will you use to determine if the manipulation is perceived as you intend? How many subjects will you recruit and from where? How long will it take? Once you have created your stimuli and manipulation check, get some subjects to take it and see if your stimuli worked. Obtain IRB approval first if required by your institution. Write up this portion of the paper.

3. Read methods sections from four studies in your area of interest and label the items that are reported. What in this chapter is included? What isn't? What special issues are reported? Identify where in the article manipulation checks are reported—in methods, results, a footnote, appendix, or online?

Suggested Readings

Creating stimuli for research:

Graesser, A. C. 1981. *Prose Comprehension Beyond the Word.* New York, NY: Springer-Verlag.

Reeves, B., and S. Geiser. 1994. "Designing Experiments That Assess Psychological Responses." In *Measuring Psychological Responses to Media*, edited by A. Lang, 165–180. Hillsdale, NJ: LEA.

Slater, M. D. 1991. "Use of Message Stimuli in Mass Communication Experiments: A Methodological Assessment and Discussion." *Journalism Quarterly* 68 (3): 412–421.

Manipulation checks:

O'Keefe, D. J. 2003. "Message Properties, Mediating States, and Manipulation Checks: Claims, Evidence, and Data Analysis in Experimental Persuasive Message Effects Research." *Communication Theory* 13 (3): 251–274.

Tao, C., and E. P. Bucy. 2007. "Conceptualizing Media Stimuli in Experimental Research: Psychological Versus Attributed-Based Definitions." *Human Communication Research* 33: 397–426.

Fixed vs. random factors:

Chang, Yu-Hsuan A., and David M. Lane. 2016. "Generalizing Across Stimuli as Well as Subjects: A Non-Mathematical Tutorial on Mixed-Effects Models." *Quantitative Methods for Psychology* 12 (3): 201–219.

Clark, H. 1973. "The Language as Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behavior* 12: 335–359. 10.1016/S0022-5371(73)80014-3.

Raaijmakers, J. G. W. 2003. "A Further Look at the 'Language-As-Fixed-Effect Fallacy.'" *Canadian Journal of Behavioral Science* 57: 141–151. 10.1037/h0087421

Raaijmakers, J. G. W., J. M. C. Schrijnemakers, and F. Gremen. 1999. "How to Deal With 'The Language-As-Fixed-Effect Fallacy': Common Misconceptions and Alternative Solutions." *Journal of Memory and Language* 41: 416–426. 10.1006/jmla.1999.2650.

For Review-No Commercial Use(2023)

Notes

1. R. Barker Bausell, *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist* (Thousand Oaks, CA: Sage, 1994), 95.
2. James J. Gibson, "The Concept of the Stimulus in Psychology," *American Psychologist* 15, no. 11 (1960): 694–703.
3. Ibid.
4. Michael D. Slater, "Messages as Experimental Stimuli: Design, Analysis, and Inference." Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication (72nd, Washington, DC, August 10–13, 1989).
5. Esther Thorson, Robert H. Wicks, and Glenn Leshner, "Experimental Methodology in Journalism and Mass Communication Research," *Journalism and Mass Communication Quarterly* 89, no. 1 (2012): 112–124.
6. S. Shyam Sundar, "Theorizing Interactivity's Effects," *Information Society* 20 (2004).
7. Shana Kushner Gadarian and Bethany Albertson, "Anxiety, Immigration, and the Search for Information," *Political Psychology* 35, no. 2 (2014): 133–164.
8. Elizabeth A. Bennion and David W. Nickerson, "The Cost of Convenience: An Experiment Showing E-Mail Outreach Decreases Voter Registration," *Political Research Quarterly* 64, no. 4 (2011): 858–869.
9. Debby Vos, "How Ordinary MPs Can Make It Into the News: A Factorial Survey Experiment with Political Journalists to Explain the Newsworthiness of MPs," *Mass Communication and Society* 19, no. 6 (2016): 738–757.
10. Theresa Beltramo et al., "The Effect of Marketing Messages and Payment Over Time on Willingness to Pay for Fuel-Efficient Cookstoves," *Journal of Economic Behavior and Organization* 118 (2015): 333–345.
11. An-Sofie Claeys and Verolien Cauberghe, "Keeping Control: The Importance of Nonverbal Expressions of Power by Organizational Spokespersons in Time of Crisis," *Journal of Communication* 64 (2014): 1160–1180.
12. Horacio Alvarez-Marinelli et al., "Computer Assisted English Language Learning in Costa Rican Elementary Schools: An Experimental Study," *Computer Assisted Language Learning* 29, no. 1 (2016): 103–126.
13. Vos, "How Ordinary MPs Can Make It Into the News."
14. Beltramo et al., "The Effect of Marketing Messages."
15. Claeys and Cauberghe, "Keeping Control."
16. Alvarez-Marinelli et al., "Computer Assisted English Language Learning."
17. Angela Min-Chia Lee, "How Fast Is Too Fast? Examining the Impact of Speed-Driven Journalism on News Production and Audience Reception" (unpublished dissertation, University of Texas, 2014); Angela M. Lee, "The Faster the Better? Examining the Effect of Live-Blogging on Audience Reception," *Journal of Applied Journalism and Media Studies* 8, no. 1 (in press).
18. Nick Lin-Hi, Jacob Horisch, and Igor Blumberg, "Does CSR Matter for Nonprofit Organizations? Testing the Link between CSR Performance and Trustworthiness in the Nonprofit Versus For-Profit Domain," *Voluntas* 26 (2015): 1944–1974.
19. André Blais, Cengiz Erisen, and Ludovic Rheault, "Strategic Voting and Coordination Problems in Proportional Systems: An Experimental Study," *Political Research Quarterly* 67, no. 2 (2014): 386–397.
20. Pazit Ben-Nun Bloom, "Disgust, Harm and Morality in Politics," *Political Psychology* 35, no. 4 (2014): 495–513.
21. Nicole LaFave, "Fake Tweets & More—Simitator," http://edtechpicks.org/2016/05/fake-tweets-more-simitator/.
22. Canterbury College, "Simitator," https://www.canterburycollege.ac.uk/teaching-learning-hub/technology/simitator/.
23. Trent R. Boulter, "Following the Familiar: Effect of Exposure and Gender on Credibility of Journalists on Twitter," (dissertation, The University of Texas at Austin), https://repositories.lib.utexas.edu/bitstream/handle/2152/62635/BOULTER-DISSERTATION-2017.pdf?sequence=1&isAllowed=y).
24. Agne Kajackaite, "If I Close My Eyes, Nobody Will Get Hurt: The Effect of Ignorance on Performance in a Real-Effort Experiment," *Journal of Economic Behavior and Organization* 116 (2015): 518–524.

For Review-No Commercial Use(2023)

25. Mina Tsay-Vogel, "Me Versus Them: Third-Person Effects Among Facebook Users," *New Media & Society* 18, no. 9 (2016): 1956–1972.
26. John C. Tedesco, "Political Information Efficacy and Internet Effects in the 2008 U.S. Presidential Election," *American Behavioral Scientist* 55, no. 6 (2011): 696–713.
27. Renita Coleman, Esther Thorson, and Lee Wilkins, "Testing the Impact of Public Health Framing and Rich Sourcing in Health News Stories," *Journal of Health Communication* 16, no. 9 (2011): 941–954.
28. Sophie Lecheler and Claes H. de Vreese, "What a Difference a Day Makes? The Effects of Repetitive and Competitive News Framing over Time," *Communication Research* 40, no. 2 (2013): 147–175.
29. Claeys and Cauberghe, "Keeping Control."
30. For an example, see H. Denis Wu and Renita Coleman, "The Affective Effect on Political Judgment: Comparing the Influences of Candidate Attributes and Issue Congruence," *Journalism and Mass Communication Quarterly* 91, no. 3 (2014): 530–543.
31. Vos, "How Ordinary MPs Can Make It Into the News."
32. Claeys and Cauberghe, "Keeping Control."
33. Vos, "How Ordinary MPs Can Make It Into the News."
34. Ibid., 752.
35. Bausell, *Conducting Meaningful Experiments*, 76.
36. Ibid., 95.
37. C. Sansone, C. C. Morf, and A. T. Panter, *The Sage Handbook of Methods in Social Psychology* (Thousand Oaks, CA: Sage Publications, 2008).
38. Diana C. Mutz and Robin Pemantle, "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods," *Journal of Experimental Political Science* 2, no. 2 (2016): 192–215.
39. Renita Coleman, "The Effect of Visuals on Ethical Reasoning: What's a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass Communication Quarterly* 83, no. 4 (2006): 835–850.
40. Stanley Milgram, *Obedience to Authority: An Experimental View* (New York: Harper & Row, 1974).
41. Ibid.
42. Bausell, *Conducting Meaningful Experiments*; Sansone, Morf, and Panter, *The Sage Handbook of Methods in Social Psychology*.
43. Mutz and Pemantle, "Standards for Experimental Research."
44. Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
45. Ibid.
46. Ibid.
47. Tsay-Vogel, "Me Versus Them."
48. Daniel J. O'Keefe, "Trends and Prospects in Persuasion Theory and Research," in *Persuasion: Theory & Research*, ed. Daniel J. O'Keefe (Thousand Oaks, CA: Sage, 2002).
49. Ibid.
50. Ibid.
51. S. Jackson and S. Jacobs, "Generalizing About Messages: Suggestions for Design and Analysis of Experiments, Plus Responses by James J. Bradac and Dean E. Hewes," *Human Communication Research,* 9, no. 2 (1983): 169–181.
52. Coleman, Thorson, and Wilkins, "Testing the Impact of Public Health Framing."
53. Charles M. Judd, Jacob Westfall, and David A. Kenny, "Treating Stimuli as a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem," *Journal of Personality and Social Psychology* 103, no. 1 (2012): 54–69.
54. Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism," 119–120.
55. Glenn Leshner, Paul D. Bolls, and Erika Thomas, "Scare 'Em or Disgust 'Em: The Effects of Graphic Health Promotion Messages," *Health Communication* 24 (2009): 447–458.
56. Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
57. D. J. O'Keefe, "Message Properties, Mediating States, and Manipulation Checks: Claims, Evidence, and Data Analysis in Experimental Persuasive Message Effects Research," *Communication Theory* 13, no. 3 (2003): 251–274.
58. Judd, Westfall, and Kenny, "Treating Stimuli as a Random Factor."
59. Ibid.
60. Ibid.
61. Yu-Hsuan A. Chang and David M. Lane, "Generalizing Across Stimuli as Well as Subjects: A Non-Mathematical Tutorial on Mixed-Effects Models," *Quantitative Methods for Psychology* 12, no. 3 (2016): 201–219.

For Review-No Commercial Use(2023)

62. Judd, Westfall, and Kenny, "Treating Stimuli as a Random Factor."
63. Ibid.
64. Chang and Lane, "Generalizing Across Stimuli"; Judd, Westfall, and Kenny, "Treating Stimuli as a Random Factor."
65. Chang and Lane, "Generalizing Across Stimuli."
66. Ibid.
67. Ibid.
68. Ibid.
69. Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
70. Annie Franco, Neil Malhotra, and Gabor Simonovits, "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science* 345, no. 6203 (2014): 1502–1505.
71. M. Foschi, "Hypotheses, Operationalizations, and Manipulation Checks," in *Laboratory Experiments in the Social Sciences*, ed. Murray Webster and Jane Sell (New York: Elsevier, 2007); Mutz and Pemantle, "Standards for Experimental Research"; Thorson, Wicks, and Leshner, "Experimental Methodology in Journalism."
72. Sansone, Morf, and Panter, *The Sage Handbook of Methods in Social Psychology*.
73. D. J. O'Keefe, "Message Properties, Mediating States and Manipulation Checks."
74. Sansone, Morf, and Panter, *The Sage Handbook of Methods in Social Psychology*.
75. Mutz and Pemantle, "Standards for Experimental Research."
76. Rebecca S. McEntee, Renita Coleman, and Carolyn Yaschur, "Comparing the Effects of Vivid Writing and Photographs on Moral Judgment in Public Relations," *Journalism and Mass Communication Quarterly* 94, no. 4 (2017): 1011–1130.
77. P. C. Cozby, *Methods of Behavioral Research*, 10th ed. (New York: McGraw-Hill, 2009); B. C. Perdue and J. O. Summers, "Checking the Success of Manipulations in Marketing Experiments," *Journal of Marketing Research* 23, no. 4 (1986): 317–326.
78. Mutz and Pemantle, "Standards for Experimental Research," 194.
79. Rosie Campbell and Philip Cowley, "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment," *Political Studies* 62, no. 4 (2013): 745–765.
80. O'Keefe, "Message Properties, Mediating States, and Manipulation Checks."
81. McEntee, Coleman, and Yaschur, "Comparing the Effects of Vivid Writing."
82. O'Keefe, "Message Properties, Mediating States, and Manipulation Checks."
83. C. Tao and E. P. Bucy, "Conceptualizing Media Stimuli in Experimental Research: Psychological Versus Attributed-Based Definitions," *Human Communication Research* 33 (2007): 397–426.
84. C. Tao and E. P. Bucy, "Conceptualizing Media Stimuli."
85. Ibid.
86. Amery D. Wu and Bruno D. Zumbo, "Understanding and Using Mediators and Moderators," *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement* 3 (2008).
87. Ibid.
88. James Hall and Pamela Sammons, "Mediation, Moderation & Interaction," in *Handbook of Quantitative Methods for Educational Research*, ed. Timothy Teo (Rotterdam, Netherlands: Sense, 2013).
89. Erica Scharrer, "Hypermasculinity, Aggression, and Television Violence: An Experiment," *Media Psychology* 7, no. 4 (2005): 353–376.
90. Tao and Bucy, "Conceptualizing Media Stimuli."
91. Ibid., 413–414.
92. R. M. Baron and D. A. Kenny, "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* 51, no. 6 (1986): 1173–1182; M. D. Sobel, "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models," in *Sociological Methodology*, ed. S. Leinhardt (San Francisco: Jossey-Bass, 1982), 212–290; Andrew F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach* (New York: Guilford Press, 2013).
93. Mutz and Pemantle, "Standards for Experimental Research"; B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (New York: Chapman & Hall, 1993).
94. Mutz and Pemantle, "Standards for Experimental Research," 197.
95. Ibid.
96. Ibid.

For Review-No Commercial Use(2023)

97. Ibid.
98. Ibid.
99. Jona Linde and Barbara Vis, "Do Politicians Take Risks Like the Rest of Us? An Experimental Test of Prospect Theory Under MPs," *Political Psychology* 38, no. 1 (2017): 101–117.
100. McEntee, Coleman, and Yaschur, "Comparing the Effects of Vivid Writing."
101. Mutz and Pemantle, "Standards for Experimental Research."
102. Ibid.
103. Beltramo et al., "The Effect of Marketing Messages," 336.
104. For an example, see Vos, "How Ordinary MPs Can Make It Into the News."
105. Ibid., 747.
106. Lin-Hi, Horisch, and Blumberg, "Does CSR Matter for Nonprofit Organizations?"
107. Ibid., 1955.
108. Jueman (Mandy) Zhang, Di Zhang, and T. Makana Chock, "Effects of HIV/AIDS Public Service Announcements on Attitude and Behavior: Interplay of Perceived Threat and Self-Efficacy," *Social Behavior and Personality* 42, no. 5 (2014).

For Review-No Commercial Use(2023)

INSTRUMENTS AND MEASURES

I have been struck again and again by how important measurement is to improving the human condition.

—Bill Gates

For Review-No Commercial Use(2023)

LEARNING OBJECTIVES

- Summarize the advantages and disadvantages of questionnaires as the instrument of an experiment.
- Discuss when single-item indicators are appropriate and when indexes are superior.
- Identify some unobtrusive instruments and be able to link them to what psychological or physiological processes they measure.
- Describe how observations are used in experiments and how reliability is ensured.
- Explain the levels of measurement required for independent variables (IVs) and dependent variables (DVs) in an experiment.
- Devise an instrument for an experiment that includes appropriate levels of measurement and response choices, and verify the validity of its constructs.

Creating stimuli is only half the job in preparing the materials of an experiment for launch. The other half involves creating the **instrument** that will be used to measure the responses provoked by the stimuli. The instrument differs from the stimulus in that it is the actual device used to measure a response, whereas the stimulus is the vehicle used to deliver the manipulation or treatment. Put another way, the instrument is the measurement device, tool, or procedure researchers use to collect data produced in response to the stimuli. Measurement is not only important to improving the human condition, as the opening quote says, but also to improving the chances for success in an experiment.

Instruments can be thought of in three main categories: (1) as self-reports, completed by the experimental subject, such as a questionnaire; (2) as technological equipment like B. F. Skinner's kymograph, a device that used a stopwatch to record movement; and (3) as observations recorded by a researcher—for example, behavior of children in classrooms. This chapter will discuss the pros and cons of these types of instruments, then follow with a discussion of measurement issues.

INSTRUMENTS

Questionnaires

For Review-No Commercial Use(2023)

In many social science experiments today, the instrument is frequently a questionnaire, similar to a survey, designed to measure people's self-reported responses to some manipulation or treatment. The questionnaire is the set of questions used to gather and record subjects' responses. Questionnaires can be administered in a variety of ways—on computers, with paper and pencil, in face-to-face interviews, or on the phone.[1] Ronald Fisher's experiment of the lady tasting tea used a face-to-face interview as the instrument; the lady simply told the experimenter whether the tea or milk came first.

Writing questions is a complicated art form, with many decisions about wording, content, and placement, among other issues. There are many good texts and articles about this topic, so I do not belabor it here. Rather, I recommend sources in the "Suggested Readings" list and hit the highlights of problems I commonly see with experimental questionnaires in More About box 10.1 in this chapter.

Importantly, the questionnaire of an experiment needs to capture the abstract concepts of the dependent variables, covariates, and intervening variables (mediators and moderators) in a concrete way. The actual questions asked of subjects are the applied translations of the theoretical concepts outlined in the literature review, known as the *operationalizations*. For example, one way the concept of attention has been measured is by how much

MORE ABOUT . . . BOX 10.1
Writing Questionnaires

- Do not ask about more than one thing per question. Watch for "and," as in: "Rate your attitude toward Blacks *and* Hispanics on the following scale." People may have different attitudes for each. Having them give one answer only invites error.
- Check to see that questions make sense to ordinary people. For example, "How involved were you in this dilemma?" is meant to operationalize involvement as being absorbed by something, but could be interpreted as being *personally* involved with the issue.
- Keep questions short and direct. Make them clearly written. Avoid using researcher language like "operationalize."
- Examine each question to see if it is needed and what level of detail is required. Every question just adds time to the study, which may cause subjects to lose interest or get tired, which introduces error. If income is not going to be used in the analysis, for example, then there is no need to ask about it.
- Avoid biased and leading questions. Take out adjectives and adverbs.
- Whenever possible, use questions that have been used in other studies. The methods section, an appendix, supplemental website, or cited studies should give exact wording and response choices. Cite these.
- Double check to be sure that all variables in the hypotheses are operationalized in the methods section and that a conceptual definition for each is in the literature review. If there are any "orphans," ask if that question is really needed.

For Review-No Commercial Use(2023)

subjects remember.[2] The concept of anxiety can be measured or operationalized with questions such as "How anxious did you feel?" To measure trustworthiness of a fictitious business, one study asked subjects how much they agreed with four statements: "I find it easy to trust (the business)," "I can trust (the business) completely," "(The business) is a trustworthy organization," and "I trust the management of (the business)."[3] Exact wording of questions used in studies can be found in the methods section, footnotes, appendix, or a supplemental website. Some studies cite another study that used the same questions previously, and researchers will need to read the cited studies to track down the exact wording.

Questions also can be less directly worded than using "anxious" to measure anxiety and "trust" to measure trustworthiness, as in the previous illustrations. For example, to

measure social marketing effectiveness, one experiment used questions asking subjects about their intentions to change, behavioral intentions, and willingness to communicate.[4] These three measures had all been used by others, which the researchers cited. The researchers also included a new measure of their own creation, asking subjects about their intentions to express an attitude by clicking a "like" button.

As illustrated by this marketing study, using questions—also called **items**—that have been developed and tested by other researchers should be the first choice to ensure reliability and validity. Taking existing questions and modifying them for some specific use also is frequently done. When previous studies lack a good measure, or researchers think there is another or better way to measure a concept, developing original questions, as with the "like" button example, is appropriate. These new items should be adequately tested to ensure they measure what they are supposed to.[5] More about that in the "Measurement Issues" section later in this chapter.

Single Items vs. Indexes

Using only one question—called a **single-item indicator**—should be reserved for concrete concepts such as age, education level, other demographics, and simple concepts such as overall ratings—for example, overall effectiveness of teachers as rated by students,[6] and overall job satisfaction.[7] It pays to know which variables are valid as single items, because fewer questions can shorten a study. However, studies with inappropriate single-item measures are more likely to be rejected by journal editors and reviewers because of issues with measurement validity and reliability.[8] Abstract and multidimensional concepts such as credibility, persuasion, and anxiety should be made up of multiple questions, which are used to form an **index**, as was done in the social marketing study with the four questions. An abstract concept that is measured with only one question tends to be weaker and less stable than one with multiple questions, and stronger measures are more likely to show significance if it exists. This is illustrated by a study that measured religiosity with two questions, "Would you describe yourself as extremely religious (7) to extremely nonreligious (1)?" and "Where would you place your religious beliefs?" with response choices from extremely fundamentalist (7) to extremely liberal (1).[9] When the two-question index was used in the analysis, it showed significant differences at the level of $p < .001$. After a reviewer suggested the fundamental-to-liberal question could have biased the analysis, it was removed and the analysis done with only the single-item asking about religious beliefs. That analysis produced the same conclusion as the previous one but was significant at $p < .05$ rather than $p < .001$, presumably because the single-item indicator was weaker than the original two-question index.[10] This illustrates one reason why single-item indicators are discouraged for all but the most obvious, concrete, and single-dimensional of variables.

For Review-No Commercial Use(2023)

Instead, complex theoretical concepts should be measured with an index of multiple questions. A good example is the concept of self-efficacy, which describes a person's belief in his or her ability to execute a task successfully.[11] People draw from four sources to gauge their abilities, including their own past performance, watching others, praise and criticism, and their emotional state. Thus, questions measuring self-efficacy should capture each of these four sources. Many self-efficacy indexes, which are also sometimes called *scales*, have been tailored for specific domains from education[12] to health care[13] and beyond. There are even specific indexes for pornography avoidance[14] and information retrieval on the web in China,[15] as well as general indexes for children, parents, teachers, and much more.[16] (For a guide on constructing self-efficacy indexes, see Bandura, 2006.[17]) As an example, an education self-efficacy index in one study[18] consisted of nine items previously developed by other researchers,[19] including: "I expect to do very well in this class," "I am sure that I can do an excellent job on the problems and tasks assigned for class," and "I know that I will be able to learn the material for my class."[20]

Slight wording changes to developed indexes can make a difference. For example, asking what one "can do" is different from asking what one is "able to do"[21] and what one "will do" in self-efficacy studies.[22] It is important to read articles on index development for such guidance. Another example can be found in the reasoned action model,[23] where injunctive norms are perceptions of what other people should or ought to do, and descriptive norms are perceptions of what people are actually doing; the questions are worded to ask how much someone *approves of* a certain behavior for injunctive norms, and how much someone *does* a certain behavior for descriptive norms.[24] This is not meant to discourage researchers from adapting questions or writing their own, only to point out the steps that should be taken to ensure that newly created questions adhere to the theory guiding the concepts. In fact, many concepts are better studied with measures developed especially for particular contexts,[25] and creating new questions is encouraged.

For Review-No Commercial Use(2023)

Pros and Cons

The use of questionnaires as the instrument to measure outcomes from experimental manipulations has its strengths and weaknesses. Questionnaires mainly rely on self-reports—that is, subjects answer the questions themselves, reporting on their own attitudes, opinions, and behaviors. Self-reports are fast, cheap, and easy. Subjects can do them privately, and the anonymous nature is thought to encourage honest answers. Survey experiments conducted online may reduce demand effects, as subjects are not being influenced by the presence of an experimenter.[27] Questionnaires are a much-used instrument in social science experiments, measuring all manner of things such as memory, attitudes, emotions, personality traits, behaviors, and behavioral intentions. Yet self-reports are inherently weak and can lead to inaccurate results.[28]

QUESTIONNAIRE EXAMPLE

Figure 10.1 contains an example of a questionnaire used in a study of the framing of health stories.[26] The double underlined areas were not shown to subjects but are here for the researcher's purpose. The labels represent the theoretical constructs the questions are meant to operationalize. For example, Factual Knowledge is measured with multiple-choice questions, and the number each subject got correct is the score. Attribution of Responsibility is the theoretical construct represented by the next set of questions. To ensure that the questions measured the construct accurately, a Cronbach's alpha measure of reliability for the purpose of internal consistency was computed; that is reported as Alpha = .805. The "R" in double underline indicates a question that needs to be reverse coded. In other words, some questions are worded negatively and some positively; reverse coding allows for responses of 7 to indicate positive responses and 1 to indicate negative.

FIGURE 10.1 ⬢ QUESTIONNAIRE

Please respond to the following items about the immigrant health story.

FACTUAL KNOWLEDGE; correct answer is in italics

What is the fastest growing segment of the U.S. population?

(a) Asian immigrants *(b) Latino immigrants* (c) European immigrants

For Review-No Commercial Use(2023)

What percentage of foreign-born adults lacks a primary health care provider?

(a) 1 in 4 (b) 1 in 3 (c) 1 in 5

What percent of immigrant workers in Santa Clara County are uninsured?

(a) 37% *(b) 54%* (c) 65%

Latinos are how much more likely to die from prostate cancer than Caucasians?

(a) four times as likely *(b) twice as likely* (c) three times as likely

Please indicate how much you agree with the following statements:

ATTRIBUTION OF RESPONSIBILITY ALPHA = .805

	Strongly Disagree						Strongly Agree
The government should spend more money on health care prevention and treatment for immigrants	1	2	3	4	5	6	7
Immigrant health care is a societal problem	1	2	3	4	5	6	7
R Immigrant health care is an individual problem	1	2	3	4	5	6	7

	Strongly Disagree						Strongly Agree
R Immigrants have only themselves to blame for lack of health care	1	2	3	4	5	6	7
R If people work hard they can almost always find a way to obtain health care	1	2	3	4	5	6	7
Employers are to blame for providing unsafe working conditions for immigrants	1	2	3	4	5	6	7
Health care providers are partly to blame for not learning about the culture of immigrants they serve	1	2	3	4	5	6	7
Communities can help immigrants by offering free English language lessons focused on health issues	1	2	3	4	5	6	7
Employers are to blame for not providing insurance for immigrant workers	1	2	3	4	5	6	7
Lack of access to health insurance makes it harder for immigrants to manage their health	1	2	3	4	5	6	7
R Poor access to health care is something that people voluntarily put themselves at risk for	1	2	3	4	5	6	7
People who have a hard time getting health care are innocent victims	1	2	3	4	5	6	7

For Review-No Commercial Use(2023)

There are many reasons for inaccuracies in self-reports. People may be managing their own self-image or responding with socially desirable answers.[29] For example, few people will admit to holding racial prejudices. Self-reports have attempted to overcome this by asking people not what *they* think but what they think others think (e.g., "However, *we are not asking for your opinions on these matters.* In this study, we are interested in *how the opinion items are perceived*.").[30] However, self-reports are notorious for being unable to overcome the social desirability bias or impression management; thus, other instruments should be considered. Additionally, people may be reluctant to answer questions honestly about illegal behavior or sensitive questions with moral overtones.[31]

Another reason why questionnaires that rely on self-reports may produce incorrect answers is that some things may be beyond a person's ability to introspectively assess.[32]

For example, some feelings are inaccessible to people, as are some of their own personality traits.[33] In these cases, people are not intentionally answering incorrectly but are simply not even aware of their true feelings or traits. When this is the case, other instruments can produce different results. An example is research that shows conservatives are happier than liberals when measured with self-reports, but that liberals show happier behavior when unobtrusive measures are used, such as observing smiling and use of positive emotional language.[34]

In addition, people can be mistaken, misremember, exaggerate, or minimize their answers to self-reports. Another explanation may be misunderstanding of a question; with survey experiments given online, subjects cannot ask a researcher for clarification of questions they may not understand. Similarly, people may interpret the meaning of questions differently than researchers intended. I once measured involvement by asking several standard questions, including, "How involved were you with this story?" with the question intended to measure how much it engaged or interested them. The graduate students used in a pilot test of the questionnaire responded the way that was intended; the undergraduates that were the subjects of the actual study wrote notes or said during debriefing, "I'm not involved with (drugs/prostitution/homelessness/elder abuse) at all!" That question was eliminated from the index. Along the same lines, some people interpret response scales differently, with some consistently giving the most extreme ratings (1s and 7s, for example) and others sticking to the middle (3s and 4s, for example).[35]

For Review-No Commercial Use(2023)

Validity studies are often conducted to compare self-reports to other instruments. For example, questions asking about criminal offenses committed can be compared against conviction rates in criminology research. Education research is frequently concerned with the accuracy of students' self-reports; for example, a study that compared high school transcripts of courses in music with self-reports of music participation found students' self-reported participation with music ensembles to be quite accurate, but students typically overreported the amount of time they spent practicing.[36] Teachers' reports of practice time was not much more accurate.[37]

On a final note, one tip for success in experiments is that **open-ended response** options in questionnaires can prove quite enlightening and helpful. These are simply questions that have no set answers or scales for subjects to choose from but allow them to answer in their own words—for example, questions such as "Tell me about . . ." or "Why did you choose this answer?" Comments from subjects as examples to illustrate statistical findings can make results "come alive." And they can be useful for backing up claims. For example, a reviewer once asked if the subjects in a study had met the criteria of carefully

scrutinizing all sides of an issue rather than merely trying to bolster their own preconceived position in a decision-making experiment; being able to pull examples from the open-ended responses to demonstrate that they had was extremely valuable.[38] There is a rich literature in the analysis of qualitative data, but these open-ended responses also can be categorized, counted, and measured to turn qualitative data into quantitative. They do need to be categorized by a human, however (see the "Observations" section in this chapter), so are not as easy to use as closed-ended questions.

For more tips on creating successful instruments and measurements in experiments, see More About box 10.2.

Some disciplines are more inclined to use self-reports than others; economics, for example, prefers observations of behavior[39] or performance-based measures.[40] Those types of instruments and other alternatives to self-reports are covered next.

Unobtrusive and Nonreactive Measures

In addition to the many weaknesses of self-reports described earlier, they are also subject to reactivity and demand characteristics, a concept introduced in chapter 2 where it was explained that bias is introduced when a subject changes his or her responses as a reaction to being studied. To overcome these biases, instruments have been devised that are nonreactive.[46] Nonreactive instruments are those that ensure the subject is not aware of what is being tested so that the act of being measured does not influence the responses. Another term often used to mean the same thing is *unobtrusive*. Unobtrusive and nonreactive measures exist on a continuum;[47] in the most unobtrusive studies, subjects are unaware they are in a study, and even the researchers do not know the hypotheses. More commonly, subjects know they are participating in a study but do not know the aims of the research so are unable to influence their responses for or against the hypothesis. The psychological instruments described in the next section are nonreactive in the sense that subjects are unable to distort their responses, but they definitely know they are participating in a study.

For Review-No Commercial Use(2023)

Technological Instruments

Another way to measure the outcomes or responses to experimental treatments is with instruments of the technological tool, mechanical, or equipment variety. Some researchers also refer to these as *devices, hardware,* and *software*.[48] B. F. Skinner invented his own, such as the "cumulative recorder" that used a pen and roll of paper to automatically record when a rat pressed a lever or a pigeon pecked a key.[49] While computer-administered

MORE ABOUT. . . BOX 10.2
Successful Measurements

- Always give the measurement immediately after a subject receives a treatment stimulus. For example, subjects read one story and immediately answer questions or list all thoughts about it before reading the next story, then answering questions or listing thoughts about it. Do not give all questions or all thought listings after all stories have been read.
- Employ multiple measures of the dependent variable—for example, a feeling thermometer, six trait evaluations, and two agree–disagree statements.[41] This offers comparisons of the measures and reduces the chance for any particular instrument to drive results. One measure may be better than the other.
- Another approach to foiling the social desirability bias is to introduce irrelevant questions or tasks so that subjects will not be able to easily guess what the study is about. Other terms for this include intervening items, buffer items, and distractor tasks. For an example, see Emily Thorson's study on belief echoes.[42]
- Remember to reverse code items that are worded the opposite way of the majority of questions. For example, 1 should become 7, 2 becomes 6, etc.
- Data from experimental subjects does not have to be collected all at once. Many experiments, particularly those using field experiments, are conducted over a period of time. Space and equipment constraints are among the most prevalent reasons. For example, a researcher showing subjects a TV show in a laboratory setting would not likely have a large-enough room; many of the psychological and physiological instruments are expensive, and only one or a few are available, so subjects are run one at a time. Other reasons for running the experiment over time include the need to include additional treatments after learning the results from one experiment.[43] However, researchers should be cognizant of threats to validity (see chapter 5) such as events that may bias the results (e.g., history).
- Include covariates judiciously. The purpose of a covariate is to reduce variance caused by nuisance variables that are related to the dependent variable. These need to be planned in advance and measured before the treatment is given if the treatment might affect the covariate. When done correctly, covariates can increase the power of an experiment. When too many variables that are not predictors of the DV are included, they can actually reduce the power. Covariates should not be used because they correlated with the independent variables, but with dependent variables.[44]
- Demographics need only be included as covariates if they are strong predictors of the dependent variable.[45] Only the demographics that are necessary to describe the sample or are to be used in the analysis need to be included in the questionnaire. For example, questions about income can be particularly sensitive for subjects; unless income is necessary for the analysis, it may be better to not ask about it.

For Review-No Commercial Use (2023)

questionnaires also require technology, they are different from the high-tech instruments in the following section, such as functional magnetic resonance imaging (fMRI) machines that detect changes in blood flow to the brain, allowing researchers to see what parts of the brain are being activated; eye-tracking devices that record eye movement; heart rate monitors that measure attention, arousal, emotion, and effort;[50] and galvanometric skin response receptors that measure changes in the skin, like sweating, as a proxy for emotion, arousal, and attention.[51] Many of these instruments today do consist of technology and software; for example, latency response, or how fast subjects respond to a question that measures cognitive processing, is done with computer software that measures response rates in milliseconds, and web tracking software that records the links a person clicks on and how long they spend on a certain site. Here these are classified separately from questionnaires that are recorded with software because subjects are not self-reporting their responses.

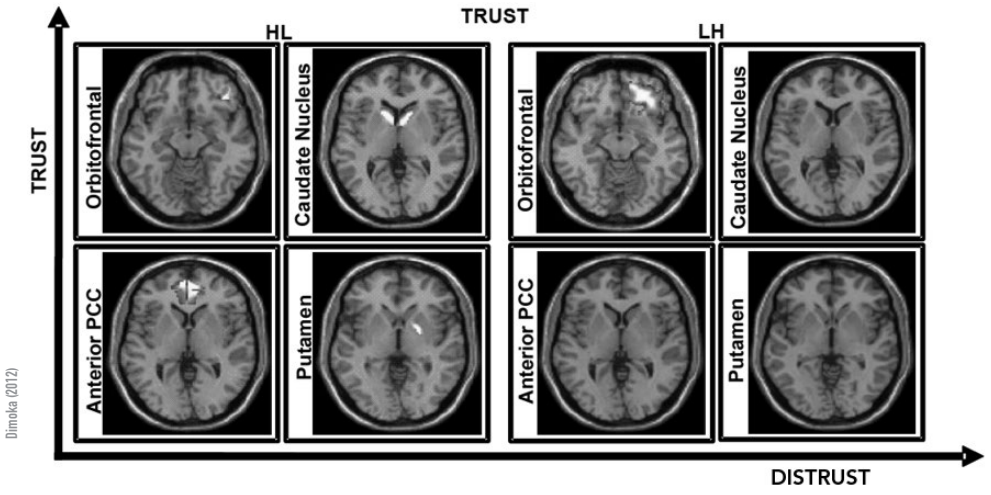
These and other instruments not included here, for the list is too long, represent **implicit tests**—that is, ones where the subject is not necessarily consciously aware of his or her responses. Many of these record psychological or physical responses such as heart rate. Some of these are used in social sciences more frequently than others, and some tend to be used in specific disciplines within the social sciences. Next are highlights of some of the more popular of these instruments in the social sciences. Space prohibits a full discussion of the strengths and weaknesses of each; this is instead an introduction, and interested readers should seek more comprehensive texts on the problems and how to address them.

For Review-No Commercial Use(2023)

Functional Magnetic Resonance Imaging

One of the newest instruments to catch on in social science is functional magnetic resonance imaging, or **fMRI**, a noninvasive brain scanner that produces cross-sectional images of the brain that captures brain activity as opposed to static brain structure. It is used to identify areas of the brain associated with certain mental processes such as trusting, or when experiencing emotions such as fear.[52] There is a growing interest in fMRI and similar neuroimaging tools across the social sciences,[53] becoming especially popular in marketing and consumer research, with many books on the

Dimoka (2012)



topic.[54] Marketing studies that use fMRI have included research on cognitive processing when people see their favorite brand label,[55] or choose between brands of beer and coffee,[56] to understand willingness to pay,[57] to see the effect of expert statements,[58] and to forecast sales changes based on point-of-sale data.[59] In political science, fMRI has been used to predict changes in attitudes toward candidates[60] and to examine the relationship between emotion and voting,[61] among other things. In health communication, fMRI has been used to examine memory for health messages[62] and to test the effectiveness of public service announcements (PSAs) about risky alcohol use.[63] In economics, it has been used to examine reputation and trust in exchanges.[64] This is not an exhaustive list of all fMRI studies in social science; nearly every discipline has at least some, including one that used fMRI to see the differences in cognitive processing between artistic and popular music.[65]

For Review-No Commercial Use(2023)

Using fMRI as an instrument is not as simple as using an instrument like a questionnaire. It takes a considerable amount of training,[66] time, and effort to conduct such a study.[67] It may even require a trained lab technician and possibly even a radiologist to look for brain abnormalities if required by an Institutional Review Board.[68] The equipment is expensive and large, consisting of a full-body tube where subjects lay down flat inside.[69] Subjects can view stimuli through goggles or on a screen, and respond verbally or on a keyboard.[70] Their heads must stay perfectly still, and they cannot have any metal piercings, implants, or medical issues.[71] The scanner is noisy, and experimental tasks need to be limited to forty-five to sixty seconds.[72] Because of the high cost and time it takes to run subjects one at a time, sample sizes are small, usually between ten and twenty subjects per study.[73] Currently, fMRI in the social sciences is used mainly to complement other instruments, allowing for triangulation with self-reports or interviews, for example, or to measure concepts subject to the biases discussed earlier.[74]

Heart Rate

Another instrument that measures physiological responses to stimuli is a heart rate monitor. Compared to fMRI, it is relatively cheap and easy to use, although not without the need for special safety training as it involves electricity.[75] Heart rate, measured via electrocardiogram, abbreviated as ECG or EKG, is used to assess attention, arousal, cognitive effort, and emotion.[76] It affords a real-time record of a person's emotional and cognitive changes while they are performing some task,[77] such as watching a political ad or listening to music. EKG measures the time between heartbeats using electrical impulses via electrodes attached to a subject's skin and requires special hardware and software.[78] It has been used in social science, including education to understand test anxiety,[79] in music to examine the differences in listening to live versus recorded music,[80] in marketing to test different breakfast drinks,[81] with public service advertising,[82] and for health messages,[83] while subjects watched jerky motion in TV and film[84] as well as commercials with celebrities,[85] while subjects read the news,[86] looked at political ads,[87] saw political leaders making inappropriate faces,[88] and with children using information technology, among other things.[89]

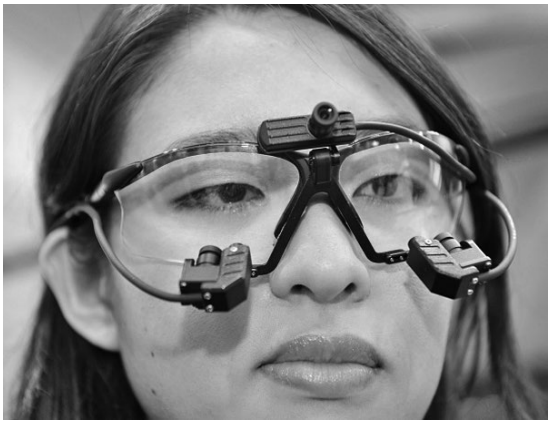
For Review-No Commercial Use(2023)

Skin Conductance

Changes in the skin also can provide measures of emotion, arousal, and attention.[90] Known by a variety of terms, including galvanic skin response (GSR) and electrodermal response (EDR), these instruments measure fluctuations in the skin including sweating, voltage, and resistance.[91] **Skin conductance** instruments are especially useful for commercial advertising, sales and marketing, and consumer research to overcome issues with self-reports.[92] In addition to these fields, skin conductance has been used in political science[93] to study the effect of anxiety on political beliefs, for example,[94] and with media violence,[95] television dramas,[96] health messages,[97] marketing,[98] teachers' stress,[99] and gaming.[100] Skin conductance instruments involve attaching electrodes to subjects' fingers or palms.[101] As with most of these high-tech instruments, subjects must be run individually, and computer or software malfunctions can further erode a small sample.[102]

Other Physiological Instruments

In addition to those already mentioned, there are instruments that measure virtually every psychological and physiological reaction a human can make. These include acoustic startle reflexes activated by loud noises,[103] saliva tests that measure stress, and voice frequency analysis and voice pitch instruments used to measure dominance,[104] among others. One health communication study used a CPR mannequin



Yoshikazu Tsuno/Staff

that recorded measurements of the quality of the compressions given, including rate, depth, complete release, and the length of time the subjects' hands were off the mannequin.[105]

Eye Tracking

Unlike the more difficult and costly instruments such as fMRI, eye-tracking devices are relatively cheap and easy to use. Eye tracking is used to unobtrusively capture a subject's visual attention patterns and direction.[106] Eye-tracking instruments measure eye fixation or position (where the eye rests), fixation frequency (how often), duration (how long),[107] and movement.[108] It is used to provide information on what a person looks at, in what order, and for how long.[109] It has been used to explain cognitive load effects[110] and to measure arousal using pupil dilation.[111]

Eye-tracking equipment used in social science experiments include devices made by companies such as SensoMotoric Instruments,[112] Tobii Technology,[113] iView Systems,[114] Applied Science Laboratories,[115] and EyeTribe,[116] among others. The cost can be as low as $200 for a unit.[117] The tracking devices are embedded in a headset[118] or glasses[119] that subjects wear, at the bottom of a mobile device or computer screen,[120] or mounted above the device.[121] Infrared cameras can be used to identify the location of the corneal reflection.[122]

For Review-No Commercial Use(2023)

Finger touches on the screen, such as zooming and swiping, also can be measured with these devices.[123] In addition to data about the eye movement and fixations, a heat map of the most looked-at areas of the stimuli can be generated,[124] among other graphics.[125]

Drawbacks of the eye-tracking instruments are some of the same as with fMRI—sample sizes are small, around twenty,[126] because the time required to run each subject is intensive.[127] For example, calibrating the device for each subject takes about ten minutes per person. Incorrectly calibrating the devices can result in a loss of subjects.[128]

Eye-tracking instruments have been used to study all manner of subjects in the social sciences, including the process of decision making;[129] news photographs;[130] the nonverbal behavior of political candidates;[131] impulse buying;[132] consumer behavior when shopping

on mobile devices;[133] advertising;[134] television;[135] economics;[136] education,[137] including how long teachers looked at students in a study where teachers wore glasses with eye-tracking devices;[138] and all things online.[139]

Eye tracking has been used in conjunction with other instruments, such as electroencephalogram, for triangulation purposes,[140] and also to validate other instruments—for example, to identify problems with questionnaires.[141] One study used eye tracking with face-to-face interviews as a measure of visual attention to study how asking different forms of questions affects answers.[142]

Eyes on Screen

A type of measure similar to eye tracking but specific to television is "eyes on screen."[143] This instrument is used to measure attention, capturing how much a subject looks at the TV screen, for how long, when, and how often the looking occurs, also known as "orienting." It requires a video camera behind a one-way mirror to record subjects as they watch TV. It is used to study TV programs and commercials, and often as a complement to or substitute for self-reports.[144] Unlike eye tracking devices which automatically collect data about where a subject looks, eyes on screen requires a human to review the video in order to determine when, how often, and how long the subjects' eyes are on the TV screen.[145] More about that under the "Observations" section.

For Review-No Commercial Use(2023)

Web Tracking

Another instrument similar to eye tracking is web tracking, which records a subject's every click and the time spent on various parts of websites. Web tracking or monitoring software has been used in various social science disciplines, including with online news use,[146] website design,[147] and political information.[148] Web monitoring software allows researchers—with subjects' informed consent—to see their actual online behavior, including when, how, and where they enter websites, their navigation paths,[149] and the queries they use in searches.[150] Advantages of this type of instrument are that users are unaware they are being observed, it is not subject to lapses in memory, and it records real-world behavior in real time.[151] Data can be collected in the lab or in subjects' own natural environments. Some researchers design their own software,[152] while others use commercially available software[153] such as Camtasia,[154] and still others analyze the log files[155] or web browser histories.[156]

Latency Response

One time-tested implicit measure that has been adopted from psychology is the latency response or reaction time instrument. This measures how quickly people respond to a question or stimulus to implicitly gauge memory, implicit attitudes, attitude accessibility, and cognitive processing.[157] It is measured in milliseconds as the length of time it takes a subject to remember something,[158] connect a stimulus to a preconceived attitude, or process new information.[159]

Latency response is well known for its use measuring attitudes about sensitive subjects, such as racial prejudice,[160] with tests such as the Implicit Associations Test (IAT).[161] These tests are designed to capture people's thoughts, attitudes, and feelings that they may be reluctant to admit, such as stereotypes.[162] They are called "implicit" because subjects are not aware of what is being measured the way they are with overt questions. The IAT, for example, shows subjects words paired with faces and asks if the word is positive or negative. Subjects who dislike certain faces, such as those of minorities, will respond significantly faster to negative words paired with them than to positive words.[163] The facial stimuli are essentially subliminal, being shown to subjects for only milliseconds, a time so short that people are not even conscious of seeing them.[164] But subconsciously, the faces of people they dislike are working to slow down their response time when the tone of the word is incongruent; that is, it takes longer to overcome a negative reaction to a disliked face and decide that the word is positive. An implicit association test is available online at Harvard's Project Implicit, https://implicit.harvard.edu/implicit/takeatest.html. Latency response also has been adapted for use in social science studies of sensitive subjects.

For Review-No Commercial Use(2023)

Another use for latency response in social science is to measure memory. For example, subjects might be shown something such as a TV show and later be shown a portion of it and asked to respond as quickly as possible to whether they saw it before or not.[165] Latency has been used as a measure of memory in advertising effectiveness[166] and is commonly found in communication studies.[167] It is also useful for education research[168]—for example, learning from multimedia.[169]

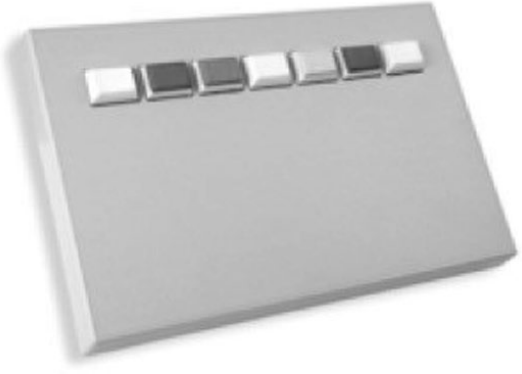


Image courtesy of Cedrus. http://cedrus.com

A related way of using latency is to measure attitude accessibility, or how automatically something is retrieved from memory, with stronger and more accessible attitudes retrieved faster than weaker attitudes.[170] For example, studies of political attitudes and behavior have used latency this way.[171]

Image courtesy of Cedrus. http://cedrus.com

Latency is also commonly used as a measure of information or cognitive processing, with longer times indicating deliberative or central route cognitive processing and shorter times indicating peripheral or heuristic processing.[172] For example, it took subjects longer to solve ethical dilemmas with photographs than without, which resulted in better moral judgment due to more careful cognitive processing.[173] Researchers also saw this phenomenon in a business study of faking on personnel tests by job applicants; creating fake answers took more mental processing and thus more time for subjects to think up than did answering honestly.[174] Similarly, in an information technology study where subjects evaluated web pages,[175] more attractive pages were linked to shorter response latencies. It took people less time to decide a page was attractive than to decide that it was not. This study also compared results from this implicit measure (latency) to explicit questions about web page attractiveness and found correspondence.

Also in this vein, latency has been used to assess persuasion knowledge, or how aware people are that they are being shown persuasive messages.[176] The theory explains that cognitive capacity is required to be aware of persuasion knowledge, and when capacity is low, people are less likely to be aware of it.[177] For example, when watching television that requires a lot of cognitive capacity, an indirect measure such as reaction time is more appropriate for measuring persuasion knowledge than is a direct measure such as asking explicit questions.

For Review-No Commercial Use(2023)

Latency responses can be measured as simply as with a researcher operating a stopwatch, or can require special equipment and software. For example, Meyer used in-person interviews where the researchers started a timer as soon as they had finished asking a question and then stopped the timer when the subject began answering.[178] Latency can also be measured over the telephone and in the laboratory with equipment consisting of a computer and software that delivers the stimulus, collects the response, and has a precision timer built in.[179] In laboratory settings or with survey experiments, computer software such as Inquisit,[180] SuperLab, or MediaLab[181] measures latency responses, reducing measurement error.[182] Within-subjects designs are preferable to control for individual differences, as some people simply read and respond faster than others. With between-subjects designs, scores should be standardized. Because of the potential for errors such as equipment malfunctions or subjects failing to respond, data should be cleaned.[183] Also, computer keyboards can be less accurate than special response devices.

Secondary Reaction Tasks

Secondary reaction task (SRT) instruments are similar to latency response in that time between a stimulus and response is measured; however, with secondary

reaction tasks, there is the addition of a "distractor" task. SRT measures attention rather than memory. In this type of study, subjects perform some task while also processing a message, effectively doing two things at one time.[184] For example, subjects may be asked to watch TV and press a button whenever they see a light flash. SRT measures cognitive processing and the limits of a person's resources to "provide clues about how much capacity is being used by the message."[185] The idea is that the increasing demands or difficulty of the first task (watching TV, reading a story, etc.) will slow down subjects' reactions to the second task (pressing a button when a tone sounds or light flashes). SRT can answer questions about attention, arousal, and involvement.[186] SRT measures the length of time in milliseconds between the time of the tone or light and when the subject presses the button. Typically, people make more mistakes on the second task or take longer to respond if the first task is more difficult or requires more effort. SRT can be important for understanding when and why people do not learn from information.

Because the measures are in milliseconds, the timing of cues must be precise and requires special equipment such as Inquisit,[187] SuperLab, or MediaLab.[188] Practice runs must be conducted to verify that the equipment works, and each subject must be familiarized with it. Multiple measures, and thus stimuli, are important to control for individual differences in baseline reaction times. For example twelve[189] to twenty[190] stimuli are typical. This also ensures against atypical messages, as discussed in the section on message variance in the last chapter. Subjects can be run a few at a time—for example, six to eight in a sitting.[191] Participants can respond on a regular computer keyboard, or with a special response pad, which may be more accurate than a keyboard. Subjects can be lost for technical reasons such as equipment failure, and for other reasons such as subjects failing to ever press the button.[192]

For Review-No Commercial Use(2023)

Continuous Response Instruments

Self-reports of attitudes, emotions, and other sensory responses reported via questionnaire have the drawback of being measured at one moment in time. But people typically experience things continuously, with their cognitive states changing. For example, when watching a TV show, emotions can range from fear to happiness to sadness and back. Questionnaires only capture the subjects' emotions at the time they answered the question. To overcome this, researchers use **continuous response instruments** that allow subjects to report their feelings or valence of thoughts and attitudes moment by moment.[193] Continuous response devices typically consist of a handheld box with a dial corresponding to 7-point scales.[194] As the subject views the stimuli, such as a TV show,

film, or advertisement, he or she can turn the dial to correspond to his or her evaluations, attitudes, emotions, or other DV. The dial hardware is connected to the TV or screen with the stimuli so that data are overlaid, showing the researcher exactly what the subject is responding to. Social science researchers use this to study PSAs,[195] facial expressions of political candidates[196] and their performance in debates,[197] and teacher effectiveness,[198] among other things.

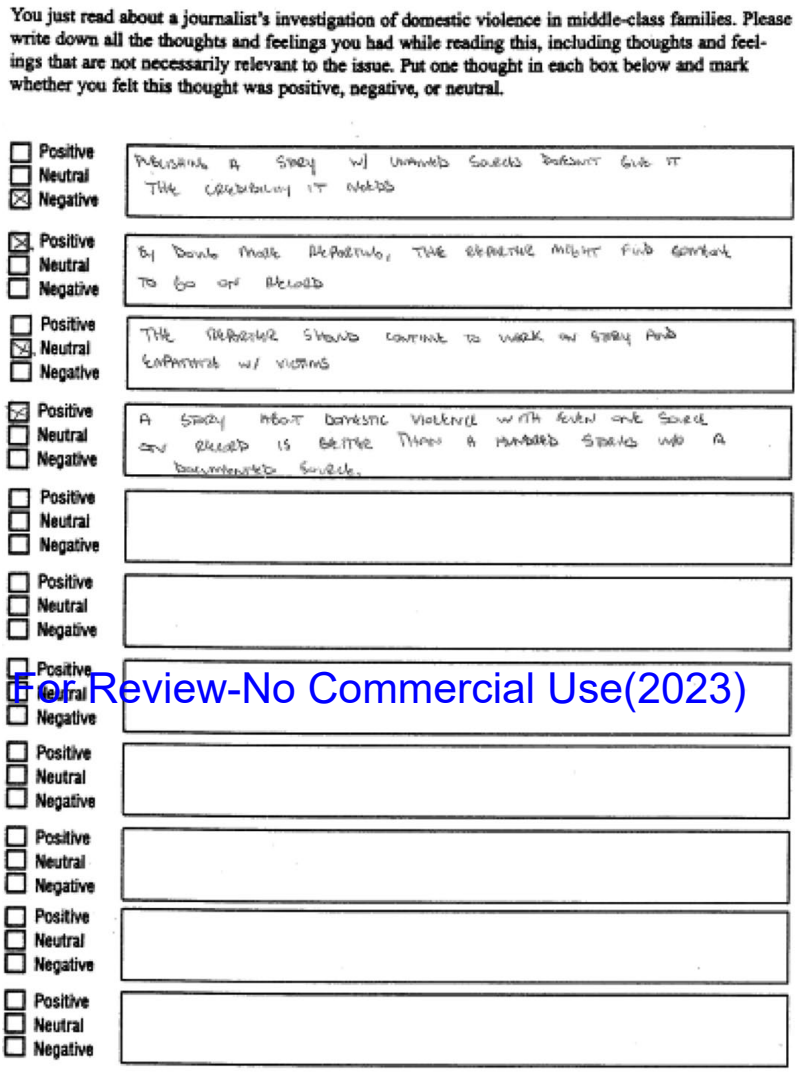
Of particular concern with all technological instruments is calibration and decay.[199] On all such instruments, buttons may become easier to push with use, creating threats to internal validity.[200] If using more than one instrument—for example, two heart rate monitors—a check should be run on the same subject at the same time to make sure that the two devices record the same value. Researchers should check the function and settings of instruments before each subject to ensure accuracy of data collected. Research assistants should be trained in the proper use of the equipment.

Thought Listing and Think Aloud

Finally, one last implicit measure that requires a simple instrument—a recording device—is the think-aloud protocol. It is discussed here with its twin measure, thought listing, which does not require any special instrument but is so similar that it is natural to discuss them together. These techniques measure how people think and solve problems and the changes in their thought processes with either a recording device (think aloud) or low-tech paper and pencil, or computer software such as those used for online survey experiments (thought listing).[201] Thought listing avoids the use of explicit questions by exposing subjects to some stimuli or having them do a task and then asking them to write down all the thoughts they had during that time. The think-aloud version accomplishes the same goal but involves having subjects say out loud everything that goes through their heads while they are being exposed to the stimuli or doing the task. Subjects are recorded and a researcher is there to prompt them to keep talking, saying everything that comes to mind. These techniques are used to investigate the mental processes that people go through for everything from using a website[202] or searching the Internet[203] to evaluating a game design[204] and making ethical decisions[205] to how people understand attempts to persuade them[206] or how they make sense of conflicting political information.[207]

For Review-No Commercial Use(2023)

These protocols avoid the problems associated with asking subjects to introspect and analyze their thought processes but do not get at the automatic processing that people are not consciously aware of.[208] The procedures can be used to measure the amount and type of


You just read about a journalist's investigation of domestic violence in middle-class families. Please write down all the thoughts and feelings you had while reading this, including thoughts and feelings that are not necessarily relevant to the issue. Put one thought in each box below and mark whether you felt this thought was positive, negative, or neutral.

☐ Positive ☐ Neutral ☒ Negative — PUBLISHING A STORY W/ UNNAMED SOURCES DOESN'T GIVE IT THE CREDIBILITY IT NEEDS

☒ Positive ☐ Neutral ☐ Negative — BY DOING MORE REPORTING, THE REPORTER MIGHT FIND COMFORT TO GO ON RECORD

☐ Positive ☒ Neutral ☐ Negative — THE REPORTER SHOULD CONTINUE TO WORK ON STORY AND EXPERIMENT W/ VICTIMS

☒ Positive ☐ Neutral ☐ Negative — A STORY ABOUT DOMESTIC VIOLENCE WITH EVEN ONE SOURCE ON RECORD IS BETTER THAN A HUNDRED STORIES W/O A DOCUMENTED SOURCE.

☐ Positive ☐ Neutral ☐ Negative

☐ Positive ☐ Neutral ☐ Negative

☐ Positive ☐ Neutral ☐ Negative

For Review-No Commercial Use(2023)

☐ Positive ☐ Neutral ☐ Negative

☐ Positive ☐ Neutral ☐ Negative

☐ Positive ☐ Neutral ☐ Negative

☐ Positive ☐ Neutral ☐ Negative

cognitive processing, whether someone thinks about the stimulus deeply and with effort or superficially, as well as the content and valence of their thoughts.

Thought listings are relatively easy to obtain, simply requiring subjects to write down all their thoughts. Researchers must count and code all the thoughts according to category or type of thought, which is covered next in the "Observations" section. One way to simplify

that task is to ask subjects to write each discrete thought in a separate box. Asking them to also check a box that indicates whether the thought was positive, negative, or neutral can take some of the guesswork out of interpreting the data. Of course, this is also subject to bias, including subjects failing to record some thoughts to make the study go faster.

Think-alouds require more time and resources to collect but overcome the issue of subjects not actually writing down every thought they had in thought listings. Using think-alouds, subjects can only be run one at a time, one per experimenter. Subjects must first be trained, as saying everything that comes to mind is not normal. Training can be done with simple math problems or word scrambles.[209] Once subjects are comfortable saying all their thoughts out loud, the stimuli are presented and the experiment begins. A researcher is positioned near the subject but out of sight to remind the subject to keep talking whenever he or she stops. It is important that the researcher not talk directly to the subjects, answer questions, or probe how or why they made some decision; researchers should only prompt them to "keep talking," or "keep saying out loud everything that comes into your head."[210] The subjects' responses should be recorded and transcribed so that, as with thought listing, they can be analyzed using content analysis methods.

Observations

For Review-No Commercial Use(2023)

Another common "instrument" in social science experiments are human observations—for example, when researchers watch students play and interact with others and record what they see.[211] Philip Zimbardo's prison experiment used researchers' observations of the subjects' behavior as they interacted as prisoners and guards. Stanley Milgram combined observations of his shock experiment subjects' comments and nonverbal behavior with the numerical data of how severe a shock the subjects were willing to give.

Observing is frequently thought of in the qualitative realm—for example, with participant observation or ethnography—but can be employed in experiments as well. Education is a field that commonly uses this technique, as it can be more reliable than asking children explicit questions they may not be able to answer. For example, in an experiment designed to assess the effectiveness of a physical education program on children's moral development, teachers observed children in physical education classes, during intramural sports, and at recess.[212] Teachers rated the students on ten prosocial behaviors, including arguing, complaining, teasing, sharing, breaking rules, and not taking turns.

Observations can be used as an instrument in other domains as well. For example, one famous study is Albert Bandura's Bobo doll experiment on violence and media content.[213] In this study, Bandura showed children TV shows with characters punching and throwing balls while shouting things like "pow" to a life-size blow-up clown that pops back

up after being punched. After children saw the TV clip, they were put in a room with a similar doll. Researchers observed through a two-way mirror and wrote down if the children imitated any of the behaviors they saw in the TV clip. In this study, none of the observers knew which children were in which treatment condition—that is, this was a double-blind study.[214] In other disciplines, observations can be used in experiments on disruptive classroom behavior, time students spend on a task, nurses' hand-washing behaviors,[215] and facial expressions,[216] among others. Experimental economics is another field that prefers observing behavior—for example, whether subjects choose a competitive or piece-rate payment for paid tasks.[217]

Interobserver Reliability

Whenever human observers are used to rate, score, judge, or code something, including data measured with instruments such as eyes on screen, thought listing, and think-alouds, reliability should be assessed to ensure that the humans doing the rating are measuring things the same way. For example, if one observer judges a child's statement as "teasing," then others should also count it as teasing, not as "arguing." This was first discussed in chapter 5 on validity and reliability. Typically, these interobserver reliability checks should involve two or more observers judging the same facial expression, behavior, language, or other element of interest, and then a calculation of agreement performed. For example, in one health communication study, observers recorded the time it took for subjects to start CPR; the observations of time were calculated for agreement by the different observers.[218] Training sessions using data not included in the actual experiment may be necessary in order to achieve acceptable levels, usually around .80.[219] (For more about interobserver reliability, see More About box 10.3.) As with technological instruments, human instruments can change over time as well—for example, by getting better at coding things accurately—and need to be checked periodically through the process.[220] There are many good guides to conducting interobserver reliability, which uses the same approach as intercoder reliability in the content analysis method. There are several recommendations in the "Suggested Readings" section.

For Review No Commercial Use (2023)

Related to training judges or coders is the issue of training research assistants who may be administering the experiment. Subjects must believe the experimental stimuli and even the situation are real, that the study is important, and find it at least somewhat interesting for good results to be obtained. Research assistants who administer the study or act as confederates—that is, they are pretending to be real subjects à la Milgram's shock study learner—must be trained to play the role, doing more acting than reading a script.[222]

MORE ABOUT . . . BOX 10.3
Interobserver Reliability

There are different statistical tools used to analyze interobserver reliability based on whether the measure is nominal, ordinal, interval, or ratio. For example, smiles can be coded as dichotomous (smiling or not smiling) or on a scale, for example 1 = frowning, 7 = smiling, 5 = neutral mouth expression, and the points in between as intermediate levels. When the measure is nominal, Scott's Pi or Cohen's Kappa are appropriate statistical tests of agreement. These calculations take into account the probability of two judges agreeing by chance rather than because they actually are interpreting things the same way. It is considered superior to Holsti's formula, which is a straight percentage of agreement that does not factor in chance. For experiments with ratio or interval level data, such as 1- to 5-point scales, Krippendorf's Alpha and Fleiss' Kappa can be used. These also allow for more than two judges.

Typical is 80% (.80) or higher agreement, although 70% (.70) is sometimes acceptable.[221]

Free online calculators for interobserver reliability exist, such as

- ReCal2, GraphPad QuickCalcs, and VassarStats for two coders (http://dfreelon.org/utils/recalfront/recal2/ and https://www.graphpad.com/quickcalcs/kappa1.cfm; http://vassarstats.net/kappa.html);
- ReCal3 for three coders (http://dfreelon.org/utils/recalfront/recal3/); and
- ReCal OIR, which calculates Krippendorf's Alpha for any number of coders (http://dfreelon.org/utils/recalfront/recal-oir/).

For Review-No Commercial Use(2023)

Others are also available. Simply search online for interobserver, interrater, or intercoder reliability calculators. In addition, there are software packages that can be purchased, including Excel, Simstat, and SPSS, which calculates Cohen's Kappa. For an overview, see http://matthewlombard.com/reliability/.

Test Name	Number of Observers	Corrects for chance?	Level of data
Holsti's formula	2	No; calculates a straight percentage	Nominal
Scott's Pi	2	Yes	Nominal
Cohen's Kappa	2	Yes	Nominal
Fleiss' Kappa	2 or more	Yes	Nominal
Krippendorf's Alpha	2 or more	Yes	All levels

Other Unobtrusive Instruments

There are many other unobtrusive instruments colloquially termed *oddball measures*.[223] For example, going to auto repair shops to see what channels radios are tuned to is an unobtrusive way to assess people's favorite radio stations. One of the more ingenious nonreactive instruments to measure attitudes is the "lost letter" technique.[224] In one classic study, stamped letters addressed to different religious groups and organizations such as Christian churches, mosques, and synagogues were scattered about a neighborhood. The address was actually a researcher's post office box, and the number of letters good Samaritans returned to a particular religious group was the measure of prejudice by people in the neighborhood. Most of these creative measures have fallen out of favor, but new approaches are taking their place. For examples such as list experiments and real effort tasks, see Study Spotlight box 10.4 for details.

Unobtrusive Questions

While this section has concentrated on nonreactive instruments other than questionnaires, it is also possible to construct questions that are unobtrusive, in the sense that subjects are unable (or less likely) to alter their responses to be socially desirable or give answers they think support the hypotheses (or social theory). The Defining Issues Test used to measure moral judgment uses such a questionnaire.[239] While subjects are aware that the study is about making ethical decisions, they are told there are no "right" answers and that the researcher is interested in what issues were most important in making the decision. The ethical dilemmas presented to subjects are written so that either course of action could be defended as ethical. Yet researchers are still able to gauge how high or low a person's moral judgment is because the twelve statements that subjects rate on a scale of importance are written to represent the six stages of moral judgment.[240] Choosing higher-staged statements results in a higher score than choosing lower-stage statements. Unless subjects are familiar with Kohlberg's theory, they are unlikely to be able to choose high-stage statements unless they have developed to that level. In fact, the instrument also includes a way for researchers to detect if subjects are trying to fake a higher score.[241] When subjects are unable to purposely alter their responses to questions, they are considered unobtrusive.

For Review-No Commercial Use(2023)

MEASUREMENT ISSUES

Measurement is another topic that is covered in greater depth in many books, textbooks, and articles, so this text does not attempt to address everything. Rather, topics of particular importance to experiments are touched on here.

STUDY SPOTLIGHT 10.4

Unobtrusive Instruments—List, Choice, and Real Effort Task Experiments

Experimentalists do not necessarily have to avoid using questionnaires in order for an instrument to be unobtrusive. Several methods have been developed that still use questionnaires but avoid asking questions directly, increasing truthful answers and avoiding nonresponses.[225] Among them are three popular with political scientists but adaptable to any social science field. In fact, the two choice-experiment studies described here come from marketing and food policy. This spotlight focuses on three types of unobtrusive instruments—the list experiment, choice experiment, and the real effort task—but there are others, such as the endorsement experiment. It should also be noted that these instruments can be used for surveys as well as experiments.

List Experiment

Kramon[226] demonstrated how a list experiment can be more externally valid in that it is more reflective of reality when the topic was the sensitive issue of vote buying. The strength of the list experiment is that it avoids the social desirability bias that can encourage participants to give socially acceptable answers rather than truthful ones. It does this by giving subjects a list of activities that are both sensitive and not sensitive, such as taking a bribe in exchange for their vote, or having seen political signs in their neighborhoods. Subjects are specifically told *not* to identify which particular items are true for them but to simply indicate *how many* items are true for them. This assumes subjects are more truthful when asked to give a number rather than answer questions directly. The control group receives only nonsensitive items, while the treatment group receives the same items as the control group but with the addition of sensitive items. The difference between group means gives the estimate of the existence of sensitive attitudes or behaviors.[227] Kramon's results showed significantly more reports of vote buying in the experiment than on surveys, which aligned with actual instances of vote buying.

For Review-No Commercial Use(2023)

Read the study at:

Kramon, Eric. 2016. "Where Is Vote Buying Effective? Evidence From a List Experiment in Kenya." *Electoral Studies* 44: 397–408. 10.1016/j.electstud.2016.09.006

Choice Experiment

In choice experiments, subjects make decisions based on trade-offs and preferences—for example, between job offers with trade-offs between salary and benefits,[228] or other types of consumer decisions including willingness to pay for certain attributes. Subjects make choices based on the characteristics of the different products or services offered, and researchers then use those to infer values the subjects hold—for example, if a higher salary is more important than a pension.

Subjects are asked to choose between alternatives rather than to rank or rate them.[229] This avoids the social desirability bias by not asking subjects what they prefer directly, but by observing the attributes of the things they choose over those they do not. For example, in a study of recycling, respondents may

(Continued)

(Continued)

self-report that they will recycle even if the packaging requires cleaning, but in a choice experiment they may actually only choose packages that do not need cleaning. That is what happened in a choice experiment that had subjects choose between two types of sandwich containers.[230] The trade-offs included how much cleaning the packages needed before recycling, how many parts it had to be separated into (paper or plastic, or both), and how long it would take to recycle. Subjects were more likely to recycle when a package had to be separated than when it had to be cleaned or took more time to recycle.

Another choice experiment on food policy found that subjects valued labels certifying the safety of beef over the country the beef came from, tenderness, or the ability to trace where foodborne diseases came from.[231] Subjects were asked to choose between two types of rib-eye steaks that came with labels indicating different prices per pound, food safety inspections, country of origin, if tenderness was guaranteed, and if it was traceable to the farm of origin.

Experts recommend choice experiments limit the number of attributes in a choice to six so as not to overwhelm subjects and make it harder for them to decide.[232] The food safety study used five attributes, for example.[233] It is also recommended that researchers use between twenty-five and ninety choice sets or scenarios in order to estimate main effects and two-way interactions.[234] The recycling study used twenty different choice scenarios, with each subject evaluating five.[235] A no-choice option is also recommended for these types of experiments because that is a realistic element that is commonly available in settings such as consumer decisions.[236]

Read these two studies at:

For Review-No Commercial Use(2023)

Klaiman, K., D. L. Ortega, and C. Garnache. 2017. "Perceived Barriers to Food Packaging Recycling: Evidence From a Choice Experiment of US Consumers." ***Food Control 73*****, 291–299.**

Loureiro, M. L., and W. J. Umberger. 2007. "A Choice Experiment Model for Beef: What U.S. Consumer Responses Tell Us About Relative Preferences for Food Safety, Country-of-Origin Labeling and Traceability." ***Food Policy*** **34 (4): 496–514.**

Real Effort Task

Another instrument that helps overcome demand effects of a laboratory setting where subjects might behave differently because researchers are watching is a real effort task experiment. In one study, researchers sought to learn the optimal frequency of reminders to increase charitable giving but wanted to avoid having subjects in a lab experiment give more to charity than they actually would if they were not being observed.[237] The study, conducted across three months, paid subjects for performing a task where they counted the 1s in a 5 x 5 matrix and received payment for correct answers in an online account. The researchers sent subjects reminders to check their online balances and told them they had the opportunity to donate to a charity via their online account. The online web portal linked to the charity's website so subjects could find out more about it. The treatment conditions were three reminder intervals: none, weekly, or monthly. The monthly reminders worked the best, and money donated was actually given to the charity.

Read this study at:

Sonntag, A., and D. J. Zizzo. 2015. "On Reminder Effects, Drop-Outs and Dominance: Evidence From an Online Experiment on Charitable Giving." ***PLoS ONE*** **10 (8): 1–17.**

A Validation Study

Another study is of special interest for this spotlight because it tested three unobtrusive methods against direct questions to determine the strengths and weaknesses of each.[238] It found that all three unobtrusive instruments—list experiments, choice experiments, and the randomized response method—were better than direct questions on sensitive issues for reducing nonresponse and providing results closer to actual behavior—in this study, votes. It found the randomized response technique performed the best but was harder for subjects to do unless they were given a practice question. All three unobtrusive instruments were significantly better than direct questions at avoiding the social desirability bias and nonresponse. Drawbacks of list experiments included ceiling and floor effects. This study also provides details on how to construct these three types of instruments.

Read this study at:

Rosenfeld, B., K. Imai, and J. N. Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." ***American Journal of Political Science*** **60 (3): 783–802**

Levels of Measure

For Review-No Commercial Use(2023)

No matter how data are collected, variables must be measured appropriately for an experiment. In the case of experiments where a test of differences such as a *t* test or analysis of variance are to be used to analyze data, the dependent variable must be measured continuously—that is, at the interval or ratio level—and the independent variables must be categorical. As the independent variable is typically the treatment and control groups, this is less of an issue than how the dependent variable is measured. Intervening variables can be measured either continuously or categorically; I typically recommend continuous measures for these as well. One reason is because continuous measures can be collapsed into categories and used as independent variables; for example, in one study, involvement was used as both an intervening variable as a continuous measure and then collapsed into high, medium, and low categories and used as an independent variable to test an interaction effect.[242] While continuous variables can be turned into categorical ones, categorical variables cannot become continuous.

Response Choices

In questionnaires, the choice of responses an experimenter gives to subjects are known as **response scales.** These choices of answers to questions range on a continuum of direction and intensity that allows researchers to assign scores to subjects. For example, a **Likert scale**[243] offers a fixed-choice response format that measures levels of agreement, frequency, importance, or likelihood—for example, Strongly

Agree (7) to Strongly Disagree (1), with the in-between points of Agree (6), Somewhat Agree (5), Neither Agree Nor Disagree (4), Somewhat Disagree, (3), and Agree (2). On a 5-point scale, the "somewhat" choices are eliminated. The wording changes to reflect the most appropriate answer to the question. For example, to "How much did you experience the following emotions," the response choices would range from Never to A Lot. As with question wording, following the lead of previous research is recommended; often, questions that have been used repeatedly and demonstrate reliability and validity will have a set response scale accompanying them. Scales are preferable to binary response options such as yes/no, true/false, or other categorical responses such as three choices of yes/no/maybe for questions about attitudes, opinions, beliefs, emotions, and other variable items. In politics especially, research has shown that scales are better at capturing people's uncertainty about their beliefs and attitudes than are dichotomous measures.[244] Categorical questions, such as one's political party, can be answered with categories, but typically, political identification is also measured with the strength of party affiliation (Would you call yourself a strong Republican/Democrat or not so strong?) and which party one feels closer to (Democrat, Republican, Neither).

At the other end of the spectrum, feeling thermometers range from 0 to 100; for example, a typical feeling thermometer from political science says, "Please rate both the candidates you read about on the thermometer below. The thermometer runs from 0 to 100 degrees. Rating above 50 means that you feel favorable and warm toward the person. Rating below 50 means that you feel unfavorable and cool toward the person."[245]

For Review-No Commercial Use(2023)

An odd number of response choices allows for a neutral point. There is debate of whether this is good or bad, and it also depends on what is being measured. For example, real attitudes may be missed with a neutral option.[246] Some questions do not need a neutral or would allow subjects with real opinions to opt out of the question. For example, "How interested are you in politics and public affairs" logically has no neutral; the four choices of very, somewhat, slightly, and not at all interested covers all bases.

Measuring things the same way as much as possible is also recommended so that they can be indexed more easily. More about indexing follows.

Ceiling and Floor Effects

Having enough response choices is important to avoid **ceiling effects**, which occur when a variable has not been measured highly enough and so scores bunch up at the top because subjects cannot respond any higher—for example, using income responses that top out at $100,000. Today, there are many more people with incomes of $100,000 or

more than there used to be, so this is no longer high enough. This results in not enough variation in the data, which means some effects on the dependent variable may not be detected. If the means of the groups are not significantly different, it could be because of ceiling effects. This can be tested by examining the distribution for normality in the curve. A similar phenomenon is called **floor effects**, when the response choices do not go low enough. Ceiling and floor effects can occur with Likert-scale responses as well as categories; for example, one study used 5-point scales and found ceiling effects.[247] Seven-point scales may have been a better choice.

Construct Validity

As was covered in chapter 5, validity is designed to test how much an instrument measures what it claims to measure. That chapter was devoted to the theoretical concept of validity; this section is about how to test it.

One of the benefits of using questions that have been used before is that they have likely undergone testing to ensure they measure what they claim to, which is called **construct validity**, first introduced in chapter 5. Other kinds of validity include face, convergent, and discriminant validity, and readers are referred to other texts in the "Suggested Readings" section for more on these. Construct validity measures how the variable represents the theoretical ideas or concepts behind it—for example, "personality," "happiness," or "depression." A test must measure what it claims to, not something similar. Construct validity is measured by expert assessment of the degree to which the instrument measures the construct, or by comparison to other instruments that measure the same or similar constructs.

For Review-No Commercial Use(2023)

Construct validity can be assessed with statistical methods that show whether there is a common factor underlying the individual items that make up the composite index or combination of several questions. To form an index, questions measured with the same response set are usually summed and averaged, or divided by the total number of questions. Before constructing the index, however, it is necessary to determine that all the questions "hang together" or are related to each other and represent the same concept. A simple way of looking at it is if the multiple questions designed to measure one concept are correlated with each other. Correlation coefficients are rarely used to determine this, however; instead, special statistics that are used to determine validity include factor analysis, among others. Cronbach's alpha is a measure of the correlation of two tests of the same construct and is used as an internal consistency estimate of reliability. These statistical tools have the advantage of helping improve the strength of an index by identifying items that should not be included.

HOW TO DO IT 10.5

Writing Up the Instrument and Measurement Sections

After the description of the stimuli (chapter 9), the methods section contains details about the instrument and measurements, and the procedures used to carry them out. All measurements for dependent variables, covariates, mediators, and moderators should give exact wording for questionnaires. Demographics are not usually given in this much detail; some articles will specify the demographics by name (e.g., age, education, gender), while others will just say "the usual demographics." If there are very many variables to operationalize, examples can be given in the text with the complete questionnaire in an appendix.[248] The response choices also should be provided—the first time in detail, and then other measures can give a briefer version. For example, that the response scale is a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5) should be explained the first time it is described, but can be more concisely explained as "the same 5-point scales" or with the endpoints in parentheses (e.g., 1 = never, 5 = all the time) on future references.

Cronbach's alpha or factor analyses for indexes and explanations of how they were formed must also be reported. Some journals use symbols, while others spell out statistical tests (e.g., Cronbach's alpha vs. the symbol α). Some journals include this in the methods section, and others put it in results; follow the style of the journal to which the paper is being submitted. The reporting of instruments and measures should be comprehensive enough that other researchers could replicate a study without having to contact the authors. Here are some examples of how authors have written about their instruments and measures. Only some of the more common social science instruments are included here; for others, see any of the studies cited in this chapter.

For Review-No Commercial Use(2023)

Example 1: Questionnaire

This experiment used a questionnaire, with the instrument items and indices' statistics in the text. Notice also how the response choices are given.

"Assessment of the current climate of opinion. We measured the participants' perceptions of the current climate of opinion using two items ("At the moment, the majority of Cologne's citizens are against the extension" [CC 1, current climate of opinion 1] and "Right now Cologne's citizens do not want an extension of the express line" [CC 2, current climate of opinion 2]). Participants could answer on a five-point Likert-type scale ranging from 1 (I don't agree at all) to 5 (I totally agree). Both items made up a scale measuring participants' perception of disagreement within the population (Spearman-Brown coefficient = .70, M = 3.75, SD = 1.21)."[249]

From: Zerback, T., T. Koch, and B. Kramer. 2015. "Thinking of Others: Effects of Implicit and Explicit Media Cues on Climate of Opinion Perceptions." ***Journalism and Mass Communication Quarterly*** **92 (2): 421–443.**

SAGE Journal Article: study.sagepub.com/coleman

Example 2: Questionnaire

This experiment also used a questionnaire, with a feeling thermometer and two other measures, with the exact wording in an appendix. It gives the reliability statistics of the index in parentheses.

"The candidate evaluation index consisted of nine variables: a feeling thermometer, six trait evaluations, and two agree-disagree statements about McKenna' s suitability for office.[6] Evaluations of McKenna formed a highly reliable index (α. = 924)."[250]

From: Thorson, Emily. 2016. "Belief Echoes: The Persistent Effects of Corrected Misinformation." ***Political Communication*** **33 (3): 460–480.**

Example 3: Construct Validity

This experiment reported more details from factor analysis and Cronbach's alpha within the text.

"To ensure construct, convergent, and discriminant validity (Perdue and Summers 1986), we conducted a confirmatory factor analysis. The items of each factor loaded only on that factor. We estimated Cronbach's (1951) alpha values to measure the reliability of any group of indicators that formed one factor. The values of all but one construct were greater than 0.70 (Nunnally 1978), in support of internal consistency; we also accepted the perceived self-efficacy construct, whose Cronbach's alpha was 0.61, which was greater than 0.60 (Robinson et al. 1991). Finally, to assess discriminant validity, we compared the average variance extracted with the squared correlations for all other constructs; the average variance extracted was always higher, indicating discriminant validity (Fornell and Larcker 1981; see Table 4 in Appendix). Thus the factor analyses revealed satisfactory results for all variables (Bagozzi et al. 1991; Cronbach et al. 1963; Nunnally 1978). We provide the constructs and results in Table 3 in the Appendix."[251]

From: Thaler, Julia, and Bernd Helmig. 2013. "Promoting Good Behavior: Does Social and Temporal Framing Make a Difference?" ***Voluntas: International Journal of Voluntary and Nonprofit Organizations*** **24 (4): 1006–1036.**

For Review, No Commercial Use(2023)

Example 4: Eye Tracking

For technological instruments, the description of the instrument and procedure can be longer than for questionnaires. Here is an example of an eye-tracking study.

"The study was conducted in individual sessions. Participants were verbally informed that the study was about visual perception. Further instructions were given on a computer screen. The instruction stated that the participants would see several pictures, and that their task would be to rate the pictures. The instructions also explained the process of the experiment to the participants: that (a) they would see a cross in the middle of the screen, (b) the cross disappears when they look at it, (c) the picture is presented for a few seconds, and (d) they would be asked for their rating.

"The presentation screen was a 22-inch monitor with a resolution of 1680 x 1050 pixels and a refresh rate of 60 Hz. Participants were seated approximately 60 cm away from the monitor. Lighting was kept constant and the participants wore headphones to control for background noises. The grey background during the task was equal to the mean luminosity of the pictures (RGB: 127,127,127).

"The experiment consisted of 40 trials. Each trial started with a fixation cross in the location where the pictures were later presented. After participants had fixated on the cross for 1,500 ms, a blank screen was presented for 500 ms. Then, one of the 40 pictures from the four categories (neutral and positive non-shopping situations, hedonic and utilitarian shopping situations) was presented on the screen for 4,000 ms. The order of the pictures was fully randomized. During the picture presentation, we assessed pupil diameter using an SMI RED 500 remote eye-tracker (Sensomotoric Instruments GmbH, Teltow, Germany) with a sampling rate of 250 Hz. After the

(Continued)

(Continued)

presentation of each picture, the participants rated how much they liked the picture (1= not at all, 7 = very much) and how positively they perceived the scene presented in the picture (i.e., valence; 1= negative, 7= positive).

"After the 40 trials, the participants answered a questionnaire that contained questions on demographics and the buying impulsiveness scale (BIS α =.83) from Rook and Fisher [1]. The BIS consists of nine items, such as 'I often buy things spontaneously.' Participants indicated on a five-point rating scale how much they agreed with each statement (1= strongly disagree, 5 = strongly agree). The authors of the BIS propose that buying impulsiveness represents a continuum; thus, the BIS does not have a cut-off point, but uses the scale score as a continuous measure of buying impulsiveness (possible range: 9–45, observed range: 9–36, M=18.3, SD=6.1)."[i]

From: Serfas, B. G., O. B. Büttner, and A. Florack. 2014. "Eyes Wide Shopped: Shopping Situations Trigger Arousal in Impulsive Buyers." ***PLoS ONE*** **9 (12): 1–9.**

Example 5: Latency Response

When psychological or physiological instruments are used, there are not typically questions that must be operationalized. In this example of a study that uses latency response, the procedure and how the instrument records the measures are described instead.

"Dependent Measure: Response Latencies. The recorded response time for participants in the lexical decision task was a measure of implicit stereotype activation. Participants received the target stimulus, which was either a word (e.g., *smart*) or a non-word (e.g., *thipy*). All non-words had lengths matching words. Participants had to decide as quickly and accurately as possible whether the target stimulus was a word or non-word by using the appropriate response keys on the computer keyboard. The first round in this task was a practice session where participants received feedback on their performances. Here, 24 target stimuli (words or non-words) appeared in quick succession in random order. An exclamation point preceded each target stimulus for 300 milliseconds. Twelve of these stimuli were neutral words (e.g., *after, complete, more*) unrelated to stereotypes. The remaining stimuli were non-words of equal length.

"Once the practice round ended, 72 target stimuli appeared in a random order. Of these, 16 stimuli were stereotypical words (e.g., *lazy, spiritual, poor*) based on pre-tests. The remaining 48 stimuli equally presented neutral words (e.g., *lady, delicious, pour*) and non-words (e.g., *linr, foigpnafe, pkid*) of length equal to the stereotypical words. Of the stereotypical words, some related to benevolent stereotypes (e.g., *traditional, polite*) and some related to hostile stereotypes (e.g., *lazy, uneducated*). An output file for each participant recorded the speed of response to each stimulus in milliseconds."[252]

From: Ramasubramanian, Srividya. 2007. "Media-Based Strategies to Reduce Racial Stereotypes Activated by News Stories." ***Journalism and Mass Communication Quarterly*** **84 (2): 249–264.**

For Review-No Commercial Use(2023)

[i]There are other types of scales, including Semantic Differential, Thurstone, and Guttman scales.

Example 6: Think Aloud and Thought Listing

As with the other unobtrusive measures, a think-aloud or thought-listing instrument describes the instructions subjects are given and procedures they follow. For example:

"Participants were then given four search tasks to complete using search engines, while following a think aloud protocol. Although they were not instructed about what search engine to use, they all selected Google's general search engine, which is popular in Iran. They were briefed about the think aloud technique (i.e. to think aloud about their decisions, actions, emotions and thoughts). The search tasks are presented below. These tasks were purposefully designed to result in some hits in search engine results that were filtered by the government so when participants clicked on some of the links they could not access the pages and were redirected to the peyvandha.ir, which indicated that the page had been blocked by the government.

"Search tasks

"Suppose for one of your course assignments or due to your personal interests you need information on the items listed below. Please open a browser and use search engines to find relevant information on the topics. Think aloud about your actions, decision, emotions and thoughts while doing the tasks:

1. You need information on the biography, beliefs and a sample of works by Mr. Shojaeddin Shafa;
2. You need some information on the thoughts and opinions of Mr. Mehdi Jami, journalist, blogger and contemporary researcher and a sample of his works;
3. You need to find the lyrics of a song named "Stay with me" from Akcent band; and
4. Find a brief history or account of socialism in Iran."[253]

For Review-No Commercial Use(2023)

From: Jamali, H. R., and P. Shahbaztabar. 2017. "The Effects of Internet Filtering on Users' Information-Seeking Behaviour and Emotions." ***Aslib Journal of Information Management*** **69 (4): 408–425.**

Regardless of whether one is creating new questions to form an index or using one from previous research, experiments are expected to validate all indexes used in the study and report it in either the methods or results sections (see How To Do It box 10.5). A pilot study—one topic in the next chapter—is a good way to test out a new construct and its index before using it in an actual experiment. The next chapter also addresses getting approval from an IRB, which is a necessary step before collecting data from any subjects. Another important topic that connects with this chapter, especially the section on unobtrusive instruments, is obtaining subjects' informed consent, protecting them from harm, and other ethical considerations concerning experiments.

Common Mistakes

- Using self-reports when unobtrusive instruments are better suited
- Using technological instruments without proper theoretical understanding or methodological training
- Using single questions to represent complex theoretical constructs; multiple items combined in an index should be used instead.
- Using new questions created by the researcher when validated items and indexes already exist

Test Your Knowledge

1. In most cases, indexes of several questions _____.
   a. Are no better than single-item indicators
   b. Are superior for measuring multilevel constructs
   c. Help shorten a study
   d. Have more issues with reliability and validity

For Review-No Commercial Use(2023)

2. Single-item indicators are _____.
   a. Devices that subjects use to indicate their response
   b. Questions that are worded directly
   c. Multiple questions combined to measure a single item
   d. Inferior for measuring multilevel constructs

3. Advantages of questionnaires are that they _____.
   a. Are fast, cheap, and easy
   b. Can be private
   c. Reduce demand effects
   d. All of these

4. Disadvantages of questionnaires are _____.
   a. Self-reports can be inaccurate
   b. They can be subject to the social desirability bias
   c. They are not good for things people may not be able to introspect about
   d. All of these

5. Functional magnetic resonance imaging, or fMRI, is a noninvasive brain scanner that is used to _____.
   a. Measure how fast subjects can answer questions
   b. Detect when a subject is lying
   c. Identify areas of the brain associated with certain mental processes such as trusting or emotions
   d. Measure how long a subject spends on a web page

6. When researchers watch students play and interact with others and record what they see, which instrument are they using?
   a. Latency response
   b. Questionnaire
   c. Observation
   d. Eyes on screen

7. In a study where teachers rate students on behaviors, including arguing, complaining, teasing, sharing, breaking rules, and not taking turns, what ensures they are rating them the same way?
   a. Manipulation checks
   b. Interobserver reliability
   c. Cronbach's alpha
   d. Factor analysis

For Review-No Commercial Use(2023)

8. In an experiment, the independent variable is measured at which level?
   a. Categorical
   b. Interval
   c. Ratio
   d. Continuous

9. At which level must the DV be measured in a typical experiment?
   a. Categorical
   b. Continuous
   c. Nominal
   d. Ordinal

10. At which level must covariates be measured?
    a. Nominal
    b. Ordinal

c. Interval

d. Any of these

Answers

1. b	4. d	7. b	9. b
2. d	5. c	8. a	10. d
3. d	6. c		

Application Exercises

1. Read any two experimental journal articles of your choice and then analyze and critique their methods sections. Identify the items from this chapter including the instruments used and procedures. For example, what are the response choices? Which constructs use indexes and which use single-item indicators? Where in the article is exact question wording reported—in the methods text, a footnote, appendix, or online? How does the study address reactivity or demand characteristics and other biases, such as social desirability, misunderstanding, and inability to introspect, associated with the instrument used? Is there enough information for you to replicate it? Is one article's methods section better written than the other? If so, why? The purpose of this assignment is to think critically about methods sections so you can do your own.

For Review-No Commercial Use(2023)

2. For the study you are working on in this book, devise an instrument and measures appropriate to test your hypotheses or research questions. Write three pages describing the instrument, including appropriate levels of measurement and response choices. Discuss how you will verify the validity of the constructs. If your main instrument is a questionnaire, write briefly about an unobtrusive instrument that could also be used in your study.

Suggested Readings

Question Writing

Tourangeau, R. 2004. "Experimental Design Considerations for Testing and Evaluating Questionnaires." In *Methods for Testing and Evaluating Survey Questionnaires,* edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. M. E. Lessler, J. Martin, and E. Singer, 209–224. Hoboken, NJ: Wiley.

Single-Item Indicators

Petrescu, Maria. 2013. "Marketing Research Using Single-Item Indicators in Structural Equation Models." *Journal of Marketing Analytics* 1 (2): 99–117.

*f*MRI

This paper gives guidelines on why and how to conduct fMRI studies.

Dimoka, Angelika. 2012. "How to Conduct a Functional Magnetic Resonance (fMRI) Study in Social Science Research." *MIS Quarterly* 36 (3): 811–840.

Eye Tracking

Jankowski, Jaroslaw, Jaroslaw Watrobski, Katarzyna Witkowska, and Pawel Ziemba. 2016. "Eye Tracking Based Experimental Evaluation of the Parameters of Online Content Affecting the Web User Behavior." In *Selected Issues in Experimental Economics,* edited by K. Nermend and M. Latuszynska, 311–332. Cham, Switzerland: Springer.

Technological Instruments

In addition to chapters on each instrument, there is a chapter on how to set up a lab for these measures.

For Review-No Commercial Use(2023)

Lang, Annie, ed. 1994. *Measuring Psychological Responses to Media*. Hillsdale, NJ: Erlbaum.

Interobserver Reliability

For a tutorial, see:

Hallgren, K. A. 2012. "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial." *Tutorials in Quantitative Methods for Psychology* 8 (1): 23–34.

More on Validity

Messick, S. 1995. "Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning." *American Psychologist* 50: 741–749.

Notes

1. Christoph Bartneck et al., "Comparing the Similarity of Responses Received from Studies in Amazon's Mechanical Turk to Studies Conducted Online and with Direct Recruitment," *PLoS ONE* 10, no. 4 (2015): 1–23.
2. Shana Kushner Gadarian and Bethany Albertson, "Anxiety, Immigration, and the Search for Information," *Political Psychology* 35, no. 2 (2014): 133–164.
3. Nick Lin-Hi, Jacob Horisch, and Igor Blumberg, "Does CSR Matter for Nonprofit Organizations? Testing the Link between CSR Performance and Trustworthiness in the Nonprofit

Versus For-Profit Domain," *Voluntas* 26 (2015): 1944–1974.

4. Julia Thaler and Bernd Helmig, "Promoting Good Behavior: Does Social and Temporal Framing Make a Difference?" *Voluntas: International Journal of Voluntary and Nonprofit Organizations* 24, no. 4 (2013): 1006–1036.
5. D. A. Dillman and C. D. Redline, "Testing Paper Self-Administered Questionnaires: Cognitive Interview and Field Test Comparisons," in *Methods for Testing and Evaluating Survey Questionnaires*, ed. S. Presser et al. (Hoboken, NJ: Wiley, 2004), 299–317; R. Tourangeau, "Experimental Design Considerations for Testing and Evaluating Questionnaires," in *Methods for Testing and Evaluating Survey Questionnaires*, ed. S. Presser et al. (Hoboken, NJ: Wiley, 2004), 209–224.
6. J. P. Wanous and M. J. Hudy, "Single-Item Reliability: A Replication and Extension," *Organizational Research Methods* 4, no. 4 (2001): 361–375.
7. J. P. Wanous and A. E. Reichers, "Estimating the Reliability of a Single-Item Measure," *Psychological Reports* 78, no. 2 (1996): 631–634.
8. C. Fuchs and A. Diamantopoulos, "Using Single-Item Measures for Construct Measurement in Management Research: Conceptual Issues and Application Guidelines," *Die Betriebswirtschaft* 69, no. 2 (2009): 195–210; Wanous and Hudy, "Single-Item Reliability."
9. Renita Coleman and Lee Wilkins, "The Moral Development of Journalists: A Comparison with Other Professions and a Model for Predicting High Quality Ethical Reasoning," *Journalism and Mass Communication Quarterly* 81, no. 3 (Autumn 2004): 511–527.
10. Ibid.
11. A. Bandura, *Self-Efficacy: The Exercise of Control* (New York: W. H. Freeman and Company, 1997); Albert Bandura, "Toward a Psychology of Human Agency," *Perspectives on Psychological Science* 1, no. 2 (2006): 164–180.
12. M. Kayhan Kurtuldu and Damla Bulut, "Development of a Self-Efficacy Scale Toward Piano Lessons," *Educational Sciences: Theory & Practice* 17, no. 3 (2017): 835–857.
13. Keiko Nanishi et al., "Determining a Cut-Off Point for Scores of the Breastfeeding Self-Efficacy Scale–Short Form: Secondary Data Analysis of an Intervention Study in Japan," *PLoS ONE* 10, no. 6 (2015): 1–12; R. C. Knibb, C. Barnes, and C. Stalker, "Parental Confidence in Managing Food Allergy: Development and Validation of the Food Allergy Self-Efficacy Scale for Parents (FASE-P)," *Clinical & Experimental Allergy* 45, no. 11 (2015): 1681–1689; Yanping Zhao et al., "Translation and Validation of a Condom Self-Efficacy Scale (CSES) Chinese Version," *AIDS Education & Prevention* 28, no. 6 (2016): 499–510; Franklin N. Glozah et al., "Assessing Alcohol Abstinence Self-Efficacy in Undergraduate Students: Psychometric Evaluation of the Alcohol Abstinence Self-Efficacy Scale," *Health & Quality of Life Outcomes* 13 (2015): 1–6.
14. Shane W. Kraus et al., "The Development and Initial Evaluation of the Pornography-Use Avoidance Self-Efficacy Scale," *Journal of Behavioral Addictions* 6, no. 3 (2017): 354–363.
15. Carole Rodon and Aline Chevalier, "Toward More Comprehensive Chinese Internet Users' Studies: Translation and Validation of the Chinese-Mandarin Version of the 8 Item Information Retrieval on the Web Self-Efficacy Scale (CH-IROWSE)," *International Journal of Human-Computer Interaction* 33, no. 10 (2017): 846–855.
16. A. Bandura, "Guide for Constructing Self-Efficacy Scales," in *Self-Efficacy Beliefs of Adolescents*, ed. Tim Urdan and Frank Pajares (Charlotte, NC: Information Age, 2006).
17. Ibid.
18. Syeda Shahida Batool, Sumaira Khursheed, and Hira Jahangir, "Academic Procrastination as a Product of Low Self-Esteem: A Mediational Role of Academic Self-Efficacy," *Pakistan Journal of Psychological Research* 32, no. 1 (Summer 2017): 195–211.
19. P. R. Pintrich and E. V. De Groot, "Motivation and Self-Regulated Learning Components of Classroom Academic Performance," *Journal of Educational Psychology* 82 (1990).
20. Batool, Khursheed, and Jahangir, "Academic Procrastination as a Product of Low Self-Esteem," 200.
21. Albert Bandura, "Much Ado Over a Faulty Conception of Perceived Self-Efficacy Grounded in Faulty Experimentation," *Journal of Social & Clinical Psychology* 26, no. 6 (2007): 641–658.

For Review-No Commercial Use (2023)

For Review-No Commercial Use(2023)

22. Bandura, "Guide for Constructing Self-Efficacy Scales," 308.
23. Martin Fishbein and Icek Ajzen, *Predicting and Changing Behavior* (New York, NY: Taylor & Francis, 2010).
24. Angela M. Lee, Renita Coleman, and Logan Molyneux, "From Thinking to Doing: Effects of Different Social Norms on Ethical Behavior in Journalism," *Journal of Media Ethics* 31, no. 2 (2016): 72–85.
25. Chang-Dae Ham, Michelle R. Nelson, and Susmita Das, "How to Measure Persuasion Knowledge," *International Journal of Advertising* 34, no. 1 (2015): 17–53.
26. Renita Coleman, Esther Thorson, and Lee Wilkins, "Testing the Impact of Public Health Framing and Rich Sourcing in Health News Stories," *Journal of Health Communication* 16, no. 9 (2011): 941–954.
27. Murray Webster and Jane Sell, *Laboratory Experiments in the Social Sciences* (Amsterdam: Elsevier, 2007).
28. Timothy Wilson, "Knowing When to Ask: Introspection and the Adaptive Unconscious," *Journal of Consciousness Studies* 10, no. 9/10 (2003): 131.
29. Rob Hoskin, "The Dangers of Self-Report," Science Brainwaves. http://www.sciencebrainwaves.com/the-dangers-of-self-report/
30. John B. McConahay, Betty B. Hardee, and Valerie Batts, "Has Racism Declined in America? It Depends on Who Is Asking and What Is Asked," *The Journal of Conflict Resolution* 25, no. 4 (1981): 563–579.
31. Cengiz Erisen, Elif Erisen, and Binnur Ozkececi-Taner, "Research Methods in Political Psychology," *Turkish Studies* 13, no. 1 (2013): 13–33.
32. A. I. Jack and A. Roepstorff, eds., *Trusting the Subject: The Use of Introspective Evidence in Cognitive Science*, 2 vols. (Exeter, UK: Imprint Academic, 2003); R. E. Nisbett and T. D. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (1977): 231–259.
33. Wilson, "Knowing When to Ask."
34. Sean P. Wojcik et al., "Conservatives Report, But Liberals Display, Greater Happiness," *Science* 347, no. 6227 (2015): 1243–1246.
35. Hoskin, "The Dangers of Self-Report."
36. C. K. Madsen, "A 30-Year Follow-up Study of Actual Applied Music Practice Versus Estimated Practice," *Journal of Research in Music Education* 52 (2004): 77–88.
37. C. C. Wang and D. W. Sogin, "Self-Reported Versus Observed Classroom Activities in Elementary General Music," *Journal of Research in Music Education* 45 (1997): 444–456.
38. Renita Coleman, "The Effect of Visuals on Ethical Reasoning: What's a Photograph Worth to Journalists Making Moral Decisions?" *Journalism and Mass Communication Quarterly* 83, no. 4 (2006): 835–850.
39. Werner Bönte, Sandro Lombardo, and Diemo Urbig, "Economics Meets Psychology: Experimental and Self-Reported Measures of Individual Competitiveness," *Personality and Individual Differences* 116 (2017): 179–185.
40. Ralph Hertwig and Andreas Ortmann, "Experimental Practices in Economics: A Methodological Challenge for Psychologists?" *Behavioral and Brain Sciences* 24, no. 3 (2001): 383.
41. Emily Thorson, "Belief Echoes: The Persistent Effects of Corrected Misinformation," *Political Communication* 33, no. 3 (2016): 460–480.
42. Ibid.
43. Rebecca B. Morton and Kenneth C. Williams, *Experimental Political Science and the Study of Causality: From Nature to the Lab* (New York: Cambridge University Press, 2010).
44. Diana C. Mutz and Robin Pemantle, "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods," *Journal of Experimental Political Science* 2, no. 2 (2016): 192–215.
45. Ibid.
46. Eugene J. Webb et al., *Unobtrusive Measures: Nonreactive Research in the Social Sciences* (Chicago: Rand McNally, 1966); Eugene J. Webb et al., *Unobtrusive Measures: Revised Edition* (Thousand Oaks, CA: Sage, 2000).
47. Immo Fritsche and Volker Linneweber, "Nonreactive (Unobtrusive) Methods," in *Handbook of Psychological Measurement—A Multimethod Perspection,* ed. M. Eid and E. Diener (Washington, DC: American Psychological Association, 2004).
48. Annie Lang, ed., *Measuring Psychological Responses to Media* (Hillsdale, NJ: Erlbaum, 1994).
49. B. F. Skinner, "Superstition in the Pigeon," *Journal of Experimental Psychology* 38 (1948): 168–172.

For Review-No Commercial Use (2023)

50. Annie Lang, "What Can the Heart Tell Us About Thinking?" in *Measuring Psychological Responses to Media*, ed. Annie Lang (Hillsdale, NJ: Erlbaum, 1994).
51. Lang, *Measuring Psychological Responses to Media*.
52. Angelika Dimoka, "How to Conduct a Functional Magnetic Resonance (fMRI) Study in Social Science Research," *MIS Quarterly* 36, no. 3 (2012): 811–840.
53. Ibid.
54. Simone Kühn, Enrique Strelow, and Jürgen Gallinat, "Multiple 'Buy Buttons' in the Brain: Forecasting Chocolate Sales at Point-of-Sale Based on Functional Brain Activation Using fMRI," *NeuroImage* 136 (2016): 122–128.
55. S. Kuhn and J. Gallinat, "The Neural Correlates of Subjective Pleasantness," *NeuroImage* 61 (2012): 289–294.
56. M. Deppe et al., "Nonlinear Responses within the Medial Prefrontal Cortex Reveal When Specific Implicit Information Influences Economic Decision Making," *Journal of NeuroImaging* 15 (2005): 171–182.
57. H. Plassmann, J. P. O'Doherty, and A. Rangel, "Appetitive and Aversive Goal Values Are Encoded in the Medial Orbitofrontal Cortex at the Time of Decision Making," *Journal of Neuroscience* 30 (2010): 10799–10808; Joshua B. Plavnick and Summer J. Ferreri, "Single-Case Experimental Designs in Educational Research: A Methodology for Causal Analyses in Teaching and Learning," *Education Psychology Review* 25 (2013): 549–569; H. Plassmann, J. O'Doherty, and A. Rangel, "Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions," *Journal of Neuroscience* 27 (2007): 9984–9988.
58. V. Klucharev, A. Smidts, and G. Fernandez, "Brain Mechanisms of Persuasion: How 'Expert Power' Modulates Memory and Attitudes," *Social Cognitive and Affective Neuroscience* 3 (2008): 353–366.
59. Kühn, Strelow, and Gallinat, "Multiple 'Buy Buttons' in the Brain."
60. J. Kato et al., "Neural Correlates of Attitude Change Following Positive and Negative Advertisements," *Frontier in Behavioral Neuroscience* 3, no. 6 (2009).
61. N. O. Rule et al., "Voting Behavior Is Reflected in Amygdala Response Across Cultures," *Social Cognitive and Affective Neuroscience* 5 (2010): 349–355; M. I. Spezio et al., "A Neural Basis for the Effect of Candidate Appearance on Election Outcomes," *Social Cognitive and Affective Neuroscience* 3 (2008): 344–352.
62. D. Seelig et al., "Low Message Sensation Health Promotion Videos Are Better Remembered and Activate Areas of the Brain Associated with Memory Encoding," *PLoS One* 9, no. e113256 (2014).
63. Martin A. Imhof et al., "How Real-Life Health Messages Engage Our Brains: Shared Processing of Effective Anti-Alcohol Videos," *Social Cognitive and Affective Neuroscience* 12, no. 7 (2017): 1188–1196.
64. B. King-Casas et al., "Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange," *Science* 308, no. 5718 (2005): 78–83.
65. Ping Huang et al., "The Difference Between Aesthetic Appreciation of Artistic and Popular Music: Evidence From an fMRI Study," *PLoS ONE* 11, no. 11 (2016): 1–14.
66. Dimoka, "How to Conduct a Functional Magnetic Resonance (fMRI) Study in Social Science Research."
67. J. C. Culham, "Functional Neuroimaging: Experimental Design and Analysis," in *Handbook of Functional Neuroimaging of Cognition*, ed. R. Cabeza and A. Kingstone (Cambridge, MA: MIT Press, 2006), 53–82; A. W. Song, S. A. Huettel, and G. McCarthy, "Functional Neuroimaging: Basic Principles of Functional MRI," in *Handbook of Functional Neuroimaging of Cognition*, ed. R. Cabeza and A. Kingstone (Cambridge, MA: MIT Press, 2006), 21–52.
68. Dimoka, "How to Conduct a Functional Magnetic Resonance (fMRI) Study in Social Science Research"
69. Ibid.
70. Ibid.
71. Ibid.
72. Ibid.
73. Ibid.
74. Ibid.
75. Lang, "What Can the Heart Tell Us About Thinking?"
76. Ibid.
77. Ibid.
78. Ibid.
79. A. Kadir ÖZer, Cemile Ekin Eremsoy, and Emel Kromer, "Physiological Measurement of the Process of Perspective Shift in the Mental Imagery of Test

For Review-No Commercial Use(2023)

Anxiety," *Electronic Journal of Social Sciences* 15, no. 58 (Summer 2016): 903–916.

80. Haruka Shoda, Mayumi Adachi, and Tomohiro Umeda, "How Live Performance Moves the Human Heart," *PLoS ONE* 11, no. 4 (2016): 1–11.
81. René A. de Wijk et al., "ANS Responses and Facial Expressions Differentiate Between the Taste of Commercial Breakfast Drinks," *PLoS ONE* 9, no. 4 (2014): 1–9.
82. Zachary Hohman et al., "A Biopsychological Model of Anti-Drug PSA Processing: Developing Effective Persuasive Messages," *Prevention Science* 18, no. 8 (2017): 1006–1016.
83. Jie Xu, "Message Sensation and Cognition Values: Factors of Competition or Integration?" *Health Communication* 30, no. 6 (2015): 589–597.
84. Himalaya Patel et al., "Receptive to Bad Reception: Jerky Motion Can Make Persuasive Messages More Effective," *Computers in Human Behavior* 32 (2014): 32–39.
85. Mareile Opwis et al., "Weird or Wired Celebrities: Effects of Celebrity Endorsers in Energy-Commercials on Psychophysiological Response Patterns," *Work Psych: Economics Conference Proceedings* (2012): 54.
86. Kevin Wise, Paul D. Bolls, and Samantha R. Schaefer, "Choosing and Reading Online News: How Available Choice Affects Cognitive Processing," *Journal of Broadcasting & Electronic Media* 52, no. 1 (2008): 69–85.
87. Zheng Wang, Alyssa C. Morey, and Jatin Srivastava, "Motivated Selective Attention During Political Ad Processing: The Dynamic Interplay Between Emotional Ad Content and Candidate Evaluation," *Communication Research* 41, no. 1 (2014): 119–156.
88. Erik Bucy and Samuel D. Bradley, "Presidential Expressions and Viewer Emotion: Counterempathic Responses to Televised Leader Displays," *Social Science Information* 43, no. 1 (2004): 59–94.
89. Sandra Ononogbu et al., "Association Between Information and Communication Technology Usage and the Quality of Sleep Among School-Aged Children During a School Week," *Sleep Disorders* (2014): 1–6.
90. Robert Hopkins and James E. Fletcher, "Electrodermal Measurement: Particularly Effective for Forecasting Message Influence on Sales Appeal," in *Measuring Psychological Responses to Media*, ed. Annie Lang (Hillsdale, NJ: Erlbaum, 1994), 113–132; Marieke G. N. Bos et al., "Psychophysiological Response Patterns to Affective Film Stimuli," *PLoS ONE* 8, no. 4 (2013): 1–8; Jonathan Renshon, Jooa Julia Lee, and Dustin Tingley, "Physiological Arousal and Political Beliefs," *Political Psychology* 36, no. 5 (2015): 569–585; Xu, "Message Sensation and Cognition Values."
91. Hopkins and Fletcher, "Electrodermal Measurement."
92. Ibid.
93. Michael Bang Petersen, Ann Giessing, and Jesper Nielsen, "Physiological Responses and Partisan Bias: Beyond Self-Reported Measures of Party Identification," *PLoS ONE* 10, no. 5 (2015): 1–10.
94. Renshon, Lee, and Tingley, "Physiological Arousal and Political Beliefs."
95. Barbara Krahé et al., "Desensitization to Media Violence: Links With Habitual Media Violence Exposure, Aggressive Cognitions, and Aggressive Behavior," *Journal of Personality and Social Psychology* 100, no. 4 (2011): 630–646.
96. Sandra Ohlsson Gershon et al., "Following the Viewers: Investigating Television Drama Engagement Through Skin Conductance Measurements," *Poetics* 64 (2017): 1–13.
97. Gabrielle Turner-Mcgrievy, Sri Kalyanaraman, and Marci K. Campbell, "Delivering Health Information Via Podcast or Web: Media Effects on Psychosocial and Physiological Responses," *Health Communication* 28, no. 2 (2013): 101–109; Chen-Bo Zhong and Julian House, "Hawthorne Revisited: Organizational Implications of the Physical Work Environment," *Research in Organizational Behavior* 32 (2012): 3–22.
98. Martin Reimann et al., "How We Relate to Brands: Psychological and Neurophysiological Insights Into Consumer–Brand Relationships," *Journal of Consumer Psychology (Elsevier Science)* 22, no. 1 (2012): 128–142; Peter Walla, Gerhard Brenner, and Monika Koller, "Objective Measures of Emotion Related to Brand Attitude: A New Way to Quantify Emotion-Related Aspects Relevant to Marketing," *PLoS ONE* 6, no. 11 (2011): 1–7.
99. Mohammed Al-Fudail and Harvey Mellar, "Investigating Teacher Stress When Using Technology," *Computers & Education* 51, no. 3 (2008): 1103–1110.

For Review-No Commercial Use(2023)

100. Guillaume Chanel, J. Matias Kivikangas, and Niklas Ravaja, "Physiological Compliance for Social Gaming Analysis: Cooperative Versus Competitive Play," *Interacting with Computers* 24, no. 4 (2012): 306–316.
101. Bos et al., "Psychophysiological Response Patterns"; Renshon, Lee, and Tingley, "Physiological Arousal and Political Beliefs."
102. Ibid.
103. Bos et al., "Psychophysiological Response Patterns."
104. Webster and Sell, *Laboratory Experiments in the Social Sciences.*
105. Hendrika Meischke et al., "Delivering 9-1-1 CPR Instructions to Limited English Proficient Callers: A Simulation Experiment," *Journal of Immigrant and Minority Health* 17, no. 4 (2015): 1049–1054.
106. Manuel Alonso Dos Santos et al., "The Influence of Image Valence on the Attention Paid to Charity Advertising," *Journal of Nonprofit & Public Sector Marketing* 29, no. 3 (2017): 346–363; Janet Hernández-Méndez and Francisco Muñoz-Leiva, "What Type of Online Advertising Is Most Effective for E-Tourism 2.0? An Eye Tracking Study Based on the Characteristics of Tourists," *Computers in Human Behavior* 50 (2015): 618–625; C. S. Longman, A. Lavric, and S. Monsell, "More Attention to Attention? An Eye-Tracking Investigation of Selection of Perceptual Attributes During a Task Switch," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39, no. 4 (2013): 1142–1151.
107. Zijian Harrison Gong and Erik P. Bucy, "When Style Obscures Substance: Visual Attention to Display Appropriateness in the 2012 Presidential Debates," *Communication Monographs* (2016): 1–24; Hernández-Méndez and Muñoz-Leiva, "What Type of Online Advertising."
108. Yi-Chun Lin, Tzu-Chien Liu, and John Sweller, "Improving the Frame Design of Computer Simulations for Learning: Determining the Primacy of the Isolated Elements or the Transient Information Effects," *Computers & Education* 88 (2015): 280–291.
109. Ibid.
110. Ibid.
111. Benjamin G. Serfas, Oliver B. Büttner, and Arnd Florack, "Eyes Wide Shopped: Shopping Situations Trigger Arousal in Impulsive Buyers," *PLoS ONE* 9, no. 12 (2014): 1–9.
112. Yunying Dong et al., "Eye-Movement Evidence of the Time-Course of Attentional Bias for Threatening Pictures in Test-Anxious Students," *Cognition & Emotion* 31, no. 4 (2017): 781–790; Serfas, Büttner, and Florack, "Eyes Wide Shopped."
113. Guillaume Hervet et al., "Is Banner Blindness Genuine? Eye Tracking Internet Text Advertising," *Applied Cognitive Psychology* 25, no. 5 (2011): 708–716; Yoon Min Hwang and Kun Chang Lee, "Using Eye Tracking to Explore Consumers' Visual Behavior According to Their Shopping Motivation in Mobile Environments," *CyberPsychology, Behavior and Social Networking* 20, no. 7 (2017): 442–447; Nicole S. Dahmen, "Photographic Framing in the Stem Cell Debate: Integrating Eye-Tracking Data for a New Dimension of Media Effects Research," *American Behavioral Scientist* 56, no. 2 (2012): 189–203; Nora A. McIntyre, Robert M. Klassen, and M. Tim Mainhard, "Are You Looking to Teach? Cultural, Temporal and Dynamic Insights into Expert Teacher Gaze," *Learning and Instruction* 49 (2017): 41–53; Cornelia Eva Neuert and Timo Lenzner, "Incorporating Eye Tracking Into Cognitive Interviews to Detect Survey Questions," *International Journal of Social Research Methodology* 19, no. 5 (2016): 501–519.
114. Marc Resnick and William Albert, "The Impact of Advertising Location and User Task on the Emergence of Banner Ad Blindness: An Eye-Tracking Study," *International Journal of Human–Computer Interaction* 30, no. 3 (2014): 206–219.
115. Gong and Bucy, "When Style Obscures Substance."
116. Alonso Dos Santos et al., "The Influence of Image Valence."
117. Ibid.
118. Olena Kaminska and Tom Foulsham, "Real-World Eye-Tracking in Face-to-Face and Web Modes," *Journal of Survey Statistics and Methodology* 2, no. 3 (2014): 343–359.
119. McIntyre, Klassen, and Mainhard, "Are You Looking to Teach?"
120. Hervet et al., "Is Banner Blindness Genuine?"
121. Hwang and Lee, "Using Eye Tracking to Explore"; Bartosz W. Wojdynski and Hyejin Bang, "Distraction Effects of Contextual Advertising on Online News Processing: An Eye-Tracking Study," *Behaviour & Information Technology* 35, no. 8 (2016): 654–664.

122. Jaroslaw Jankowski et al., "Eye Tracking Based Experimental Evaluation of the Parameters of Online Content Affecting the Web User Behaviour," in *Selected Issues in Experimental Economics*, ed. K. Nermend and M. Latuszynska (Cham, Switzerland: Springer, 2016), 311–332.
123. Hwang and Lee, "Using Eye Tracking to Explore."
124. Alonso Dos Santos et al., "The Influence of Image Valence"; Gong and Bucy, "When Style Obscures Substance"; Wojdynski and Bang, "Distraction Effects of Contextual Advertising."
125. Jankowski et al., "Eye Tracking Based Experimental Evaluation."
126. Dong et al., "Eye-Movement Evidence."
127. McIntyre, Klassen, and Mainhard, "Are You Looking to Teach?"
128. Hernández-Méndez and Muñoz-Leiva, "What Type of Online Advertising"; Hwang and Lee, "Using Eye Tracking to Explore."
129. Anastasia G. Peshkovskaya et al., "The Socialization Effect on Decision Making in the Prisoner's Dilemma Game: An Eye-Tracking Study," *PLoS ONE* 12, no. 4 (2017): 1–15.
130. Dahmen, "Photographic Framing in the Stem Cell Debate."
131. Gong and Bucy, "When Style Obscures Substance."
132. Serfas, Büttner, and Florack, "Eyes Wide Shopped."
133. Hwang and Lee, "Using Eye Tracking to Explore."
134. Alonso Dos Santos et al., "The Influence of Image Valence"; Hervet et al., "Is Banner Blindness Genuine?"; Resnick and Albert, "The Impact of Advertising Location"; Wojdynski and Bang, "Distraction Effects of Contextual Advertising."
135. Sabrina Heike Kessler and Lars Guenther, "Eyes on the Frame," *Internet Research* 27, no. 2 (2017): 303–320.
136. Jankowski et al., "Eye Tracking Based Experimental Evaluation."
137. Lin, Liu, and Sweller, "Improving the Frame Design."
138. McIntyre, Klassen, and Mainhard, "Are You Looking to Teach?"
139. Hernández-Méndez and Muñoz-Leiva, "What Type of Online Advertising."
140. Alonso Dos Santos et al., "The Influence of Image Valence."
141. Neuert and Lenzner, "Incorporating Eye Tracking Into Cognitive Interviewing"; Hervet et al., "Is Banner Blindness Genuine?"
142. Kaminska and Foulsham, "Real-World Eye-Tracking."
143. Esther Thorson, "Using Eyes on Screen as a Measure of Attention to Television," in *Measuring Psychological Responses to Media*, ed. Annie Lang (Hillsdale, NJ: Erlbaum, 1994), 65–84.
144. Ibid.
145. Ibid.
146. D. Tewksbury, "What Do Americans Really Want to Know? Tracking the Behavior of News Readers on the Internet," *Journal of Communication* 53, no. 4 (2003): 694–710; Martijn Kleppe and Marco Otte, "Analysing and Understanding News Consumption Patterns by Tracking Online User Behaviour with a Multimodal Research Design," *Digital Scholarship in the Humanities* 32 (2017): 158–170.
147. Renita Coleman et al., "Public Life and the Internet: If You Build a Better Website, Will Citizens Become Engaged?" *New Media & Society* 10, no. 2 (2008): 179–201.
148. E. Menchen-Trevino and C. Karr, "Researching Real-World Web Use with Roxy: Collecting Observational Web Data with Informed Consent," *Journal of Information Technology & Politics* 9, no. 3 (2012): 254–268; S. A. Munson, S. Y. Lee, and P. Resnick, "Encouraging Reading of Diverse Political Viewpoints with a Browser Widget," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (Cambridge, MA: AAI Press, 2013), 419–428, http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6119, accessed Sept. 25, 2015.
149. Kleppe and Otte, "Analysing and Understanding News Consumption Patterns."
150. P. Batista and M. Silva, "Mining Web Access Logs of an On-Line Newspaper," *Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems*, http://xldb.di.fc.ul.pt/xldb/publications/rpec02.pdf. (2002).
151. Kleppe and Otte, "Analysing and Understanding News Consumption Patterns."
152. Ibid.; Menchen-Trevino and Karr, "Researching Real-World Web Use."
153. O. Findahl, *The Swedes and the Internet 2009* (Gavle: World Internet Institute, 2009).

For Review-No Commercial Use(2023)

154. Coleman et al., "Public Life and the Internet."
155 Batista and Silva, "Mining Web Access Logs."
156. S. Ebersole, "Uses and Gratifications of the Web Among Students," *Journal of Computer-Mediated Communication* 6, no. 1 (2000).
157. Ham, Nelson, and Das, "How to Measure Persuasion Knowledge."
158. Glen T. Cameron and David A. Frieske, "The Time Needed to Answer: Measurement of Memory Response Latency," in *Measuring Psychological Responses to Media*, ed. A. Lang (Hillsdale, NJ: Erlbaum, 1994), 149–164.
159. Marco Meyer and Harald Schoen, "Response Latencies and Attitude-Behavior Consistency in a Direct Democratic Setting: Evidence From a Subnational Referendum in Germany," *Political Psychology* 35, no. 3 (2014): 431–440.
160. Dermot Barnes-Holmes et al., "The Implicit Relational Assessment Procedure: Exploring the Impact of Private Versus Public Contexts and the Response Latency Criterion on Pro-White and Anti-Black Stereotyping Among White Irish Individuals," *Psychological Record* 60, no. 1 (Winter 2010): 57–79.
161. A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology* 74, no. 6 (1998): 1464–1480.
162. A. G. Greenwald et al., "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity," *Journal of Personality and Social Psychology* 97 (2009): 17–41.
163. Erisen, Erisen, and Ozkececi-Taner, "Research Methods in Political Psychology."
164. Ibid.
165. Cameron and Frieske, "The Time Needed to Answer."
166. W. Jeffrey Burroughs and Richard A. Feinberg, "Using Response Latency to Assess Spokesperson Effectiveness," *Journal of Consumer Research* 14, no. 2 (1987): 295–299.
167. Cheryl Bracken, Gary Pettey, and Mu Wu, "Revisiting the Use of Secondary Task Reaction Time Measures in Telepresence Research: Exploring the Role of Immersion and Attention," *AI & Society* 29, no. 4 (2014): 533–538; Justin Robert Keene and Annie Lang, "Dynamic Motivated Processing of Emotional Trajectories in Public Service Announcements," *Communication Monographs* 83, no. 4 (2016): 468–485; Glenn Leshner et al., "When a Fear Appeal Isn't Just a Fear Appeal: The Effects of Graphic Anti-Tobacco Messages," *Journal of Broadcasting & Electronic Media* 54, no. 3 (2010): 485–507.
168. Thierry Olive et al., "Children's Cognitive Effort and Fluency in Writing: Effects of Genre and of Handwriting Automatisation," *Learning and Instruction* 19, no. 4 (2009): 299–308.
169. Cornelia Schoor, Maria Bannert, and Roland Brünken, "Role of Dual Task Design When Measuring Cognitive Load During Multimedia Learning," *Educational Technology Research and Development* 60, no. 5 (2012): 753–768.
170. Meyer and Schoen, "Response Latencies and Attitude-Behavior Consistency."
171. Ibid.
172. Ibid.
173. Coleman, "The Effect of Visuals on Ethical Reasoning."
174. Saul Fine and Merav Pirak, "Faking Fast and Slow: Within-Person Response-Time Latencies for Measuring Faking in Personnel Testing," *Journal of Business and Psychology* 31, no. 1 (2016): 51–64.
175. Noam Tractinsky et al., "Evaluating the Consistency of Immediate Aesthetic Perceptions of Web Pages," *International Journal of Human-Computer Studies* 64, no. 11 (2006): 1071–1083.
176. Ham, Nelson, and Das, "How to Measure Persuasion Knowledge."
177. Ibid.
178. Meyer and Schoen, "Response Latencies and Attitude-Behavior Consistency."
179. Cameron and Frieske, "The Time Needed to Answer."
180. Fine and Pirak, "Faking Fast and Slow."
181. Glenn Leshner, Paul D. Bolls, and Erika Thomas, "Scare 'Em or Disgust 'Em: The Effects of Graphic Health Promotion Messages," *Health Communication* 24 (2009).
182. Cameron and Frieske, "The Time Needed to Answer."
183. Ibid.
184. M. D. Basil, "Secondary Reaction-Time Measures," in *Measuring Psychological Responses to Media*, ed. A. Lang (Hillsdale, NJ: Erlbaum, 1994), 85–98.
185. Ibid., 85.

For Review-No Commercial Use(2023)

186. Ibid.
187. Bracken, Pettey, and Wu, "Revisiting the Use of Secondary Task Reaction."
188. Leshner, Bolls, and Thomas, "Scare 'Em or Disgust 'Em."
189. Bracken, Pettey, and Wu, "Revisiting the Use of Secondary Task Reaction."
190. Keene and Lang, "Dynamic Motivated Processing of Emotional Trajectories."
191. Ibid.; Schoor, Bannert, and Brünken, "Role of Dual Task Design."
192. Bracken, Pettey, and Wu, "Revisiting the Use of Secondary Task Reaction."
193. Gong and Bucy, "When Style Obscures Substance"; Frank Biocca, Prabu David, and Mark West, "Continuous Response Measurement (CRM): A Computerized Tool for Research on the Cognitive Processing of Communication Messages," in *Measuring Psychological Responses to Media*, ed. A. Lang (Hillsdale, NJ: Erlbaum, 1994), 15–64.
194. Ibid.
195. Keene and Lang, "Dynamic Motivated Processing of Emotional Trajectories."
196. Gong and Bucy, "When Style Obscures Substance."
197. Jeremy Saks et al., "Dialed In: Continuous Response Measures in Televised Political Debates and Their Effect on Viewers," *Journal of Broadcasting & Electronic Media* 60, no. 2 (2016): 231–247.
198. Jason M. Silveira, "The Perception of Pacing in a Music Appreciation Class and Its Relationship to Teacher Effectiveness and Teacher Intensity," *Journal of Research in Music Education* 62, no. 3 (2014): 302–318.
199. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963).
200. William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Belmont, CA: Wadsworth Cengage Learning, 2002).
201. M. A. Shapiro, "Think-Aloud and Thought-List Procedures in Investigating Mental Processes," in *Measuring Psychological Responses to Media Messages*, ed. Annie Lang (Hillsdale, NJ: Lawrence Erlbaum Associates, 1994), 1–14.
202. Coleman et al., "Public Life and the Internet."
203. Hamid R. Jamali and Pria Shahbaztabar, "The Effects of Internet Filtering on Users' Information-Seeking Behaviour and Emotions," *Aslib Journal of Information Management* 69, no. 4 (2017): 408–425; Pengyi Zhang and Dagobert Soergel, "Process Patterns and Conceptual Changes in Knowledge Representations During Information Seeking and Sensemaking: A Qualitative User Study," *Journal of Information Science* 42, no. 1 (2016): 59–78.
204. Heather Desurvire and Magy Seif El-Nasr, "Methods for Game User Research: Studying Player Behavior to Enhance Game Design," *IEEE Computer Graphics and Applications* 33, no. 4 (2013): 82–87.
205. Coleman, "The Effect of Visuals on Ethical Reasoning"; Carter Gibson et al., "A Qualitative Analysis of Power Differentials in Ethical Situations in Academia," *Ethics & Behavior* 24, no. 4 (2014): 311–325.
206. Ham, Nelson, and Das, "How to Measure Persuasion Knowledge"; Robert C. Sinclair, Tanya K. Lovsin, and Sean E. Moore, "Mood State, Issue Involvement, and Argument Strength on Responses to Persuasive Appeals," *Psychological Reports* 101, no. 3 (2007): 859–873.
207. Jeffrey L. Bernstein, "What Think-Aloud Protocols Can Teach Us About How People Manage Political Information." Conference presentation at the American Political Science Association Teaching and Learning Conference, San Jose, California, February 2008.
208. Shapiro, "Think-Aloud and Thought-List Procedures."
209. For examples, see Coleman et al., "Public Life and the Internet."
210. Shapiro, "Think-Aloud and Thought-List Procedures."
211. Debra Malmberg, Marjorie Charlop, and Sara Gershfeld, "A Two Experiment Treatment Comparison Study: Teaching Social Skills to Children with Autism Spectrum Disorder," *Journal of Developmental and Physical Disabilities* 27, no. 3 (2015): 375–392.
212. Sandra L. Gibbons, Vicki Ebbeck, and Maureen R. Weiss, "Fair Play for Kids: Effects on the Moral Development of Children in Physical Education," *Research Quarterly for Exercise and Sport* 66, no. 3 (1995): 247–255.
213. Albert Bandura, "Influence of Models' Reinforcement Contingencies on the Acquisition of Imitative Responses," *Journal of Personality and Social Psychology* 1, no. 6 (1965): 589–595.

For Review-No Commercial Use(2023)

214. Ibid.
215. R. Barker Bausell, *Conducting Meaningful Experiments: 40 Steps to Becoming a Scientist* (Thousand Oaks, CA: Sage, 1994).
216. Wojcik et al., "Conservatives Report."
217. Bönte, Lombardo, and Urbig, "Economics Meets Psychology."
218. Meischke et al., "Delivering 9-1-1 CPR Instructions."
219. K. A. Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial," *Tutorials in Quantitative Methods for Psychology* 8, no. 1 (2012): 23–34.
220. Shadish, Cook, and Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.*
221. Kimberly A. Neuendorf, *The Content Analysis Guidebook* (Thousand Oaks, CA: Sage, 2002), 145.
222. Webster and Sell, *Laboratory Experiments in the Social Sciences*.
223. Webb et al., *Unobtrusive Measures.*
224. Stanley Milgram, Leon Mann, and Susan Harter, "The Lost-Letter Technique: A Tool of Social Science Research," *Public Opinion Quarterly* 27, no. 3 (1965): 437–438.
225. Bryn Rosenfeld, Kosuke Imai, and Jacob N. Shapiro, "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions," *American Journal of Political Science* 60, no. 3 (2016): 783–802.
226. Eric Kramon, "Where Is Vote Buying Effective? Evidence From a List Experiment in Kenya," *Electoral Studies* 44 (2016): 397–408.
227. Rosenfeld, Imai, and Shapiro, "An Empirical Validation Study."
228. Jing Chen and Pallavi Chitturi, "Choice Experiments for Estimating Main Effects and Interactions," *Journal of Statistical Planning and Inference* 142, no. 2 (2012): 390–396.
229. Kimberly Klaiman, David L. Ortega, and Cloé Garnache, "Perceived Barriers to Food Packaging Recycling: Evidence From a Choice Experiment of US Consumers," *Food Control* 73 (2017): 291–299.
230. Ibid.
231. M. L. Loureiro and W. J. Umberger, "A Choice Experiment Model for Beef: What U.S. Consumer Responses Tell Us About Relative Preferences for Food Safety, Country-of-Origin Labeling and Traceability," *Food Policy* 32, no. 4 (2007): 496–514.
232. Chen and Chitturi, "Choice Experiments for Estimating Main Effects and Interactions."
233. Loureiro and Umberger, "A Choice Experiment Model for Beef."
234. Chen and Chitturi, "Choice Experiments for Estimating Main Effects and Interactions."
235. Klaiman, Ortega, and Garnache, "Perceived Barriers to Food Packaging Recycling."
236. Loureiro and Umberger, "A Choice Experiment Model for Beef."
237. Axel Sonntag and Daniel John Zizzo, "On Reminder Effects, Drop-Outs and Dominance: Evidence From an Online Experiment on Charitable Giving," *PLoS ONE* 10, no. 8 (August 7, 2015): 1–17.
238. Rosenfeld, Imai, and Shapiro, "An Empirical Validation Study."
239. James R. Rest et al., *Postconventional Moral Thinking: A Neo-Kohlbergian Approach* (Mahwah, NJ: Erlbaum, 1999).
240. Lawrence Kohlberg, *The Philosophy of Moral Development: Moral Stages and the Idea of Justice* (Cambridge, MA: Harper & Row, 1981); Lawrence Kohlberg, *The Psychology of Moral Development: The Nature and Validity of Moral Stages* (San Francisco: Harper & Row, 1984).
241. James R. Rest, Lynne Edwards, and Stephen J. Thoma, "Designing and Validating a Measure of Moral Judgment: Stage Preference and Stage Consistency Approaches," *Journal of Educational Psychology* 89, no. 1 (March 1997): 5–28.
242. Coleman, "The Effect of Visuals on Ethical Reasoning."
243. R. M. Alvarez and C. H. Franklin, "Uncertainty and Political Perceptions," *Journal of Politics* 56, no. 3 (1994): 671–688.
244. Emily Thorson, "Belief Echoes."
245. Erisen, Erisen, and Ozkececi-Taner, "Research Methods in Political Psychology."
246. H. Cho and F. J. Boster, "Effects of Gain Versus Loss Frame Antidrug Ads on Adolescents," *Journal of Communication* 58, no. 3 (2008): 428–446.
247. Alan Gerber et al., "Reporting Guidelines for Experimental Research: A Report From the Experimental Research Section Standards Committee," *Journal of Experimental Political Science* 1, no. 1 (2014): 81–98.

For Review-No Commercial Use (2023)

248. Thomas Zerback, Thomas Koch, and Benjamin Kramer, "Thinking of Others: Effects of Implicit and Explicit Media Cues on Climate of Opinion Perceptions," *Journalism and Mass Communication Quarterly* 92, no. 2 (2015): 421–443.
249. Thorson, "Belief Echoes," 467.
250. Thaler and Helmig, "Promoting Good Behavior," 1018.
251. Srividya Ramasubramanian, "Media-Based Strategies to Reduce Racial Stereotypes Activated by News Stories," *Journalism and Mass Communication Quarterly* 84, no. 2 (Summer 2007): 249–264.
252. Jamali and Shahbaztabar, "The Effects of Internet Filtering," 412.

For Review-No Commercial Use(2023)

For Review-No Commercial Use(2023)