Experiments for Evaluating Survey Questions

Jon A. Krosnick

Stanford University

June, 2010

Jon Krosnick is University Fellow at Resources for the Future. Correspondence regarding this manuscript should be addressed to Jon Krosnick, 432 McClatchy Hall, 450 Serra Mall, Stanford University, Stanford, CA 94305 (e-mail: Krosnick@stanford.edu).

Experiments for Evaluating Survey Questions

Introduction

For 100 years, experiments have been conducted that allow researchers to compare different ways of asking the same question in measuring a single construct. These experiments have not always been conducted with the purpose of identifying flaws in question design or practices for obtaining the most accurate measurements. But regardless of their original purpose, these experiments have accumulated into a gigantic literature with implications for best practices in questionnaire design.

In this paper, I review the experimental method for evaluating questions and identifying their flaws. I begin by reviewing the logic of experimentation and how experiments can be valuable for evaluating question functioning. I then talk about how experimental data are produces and analyzed. A discussion of the assumptions underlying the method is followed by a listing of the types of insights that can be gained from experiments, how problems with questions can be characterized, how the method can be conducted misleadingly, whether experiments re suitable to cross-cultural investigations, how other question evaluation methods can be coordinated with experimentation, and what methodological criteria should be used for including the method in Q-Bank.

The Logic of Experimentation

Treatment vs. Control Groups

Experimentation is one of the oldest methods of scientific investigation. In lay

language, to experiment with something is to test it out and to explore how it works. But in science, experimentation is a formal method for investigating causal relations. The vast majority of experiments compare what occurs under one set of circumstances with what occurs under a different set of circumstances, to assess whether the change in circumstances is responsible for any observed change in events. For example, if one wanted to assess whether taking a particular drug causes a reduction in blood cholesterol levels, one group of research participants would take the drug regularly for an extended time period (called the "treatment group"), and another group of participants would not take the drug (called the "control group"). At the end of the time period, blood cholesterol levels of both groups could be compared, to see whether the former group's levels are lower than the latter's.

Confounds

A number of important principles guide optimal experimental design to maximize a scientist's ability to reach a valid inference from the study's result. For example, it is essential to minimize confounds in the design. A confound is a difference between the two groups of participants that could be responsible for observed differences between the treatment and control groups. One potential source of confounds is participant self-selection into the treatment and control groups. Imagine that a researcher allowed each participant to decide whether to be in the treatment group or the control group of the study. And imagine that people who have higher cholesterol levels before the study tend to choose to be in the treatment group (in the hope that the drug will lower their cholesterol), whereas people with lower pre-existing cholesterol levels tend to choose to

be in the control group more often (because they have no special motivation to lower their cholesterol and want to generously allow others who need such help to be in the treatment group). If the treatment group ended the study with a higher cholesterol level than the control group, this difference might have been present even before the study began. Therefore, experiments routinely involve random assignment of participants to the treatment and control groups to eliminate the potential for systematic self-selection driven differences between the groups.

Other potential confounds in an experimental design can come from experiences that vary between the treatment and control groups that are not the intended treatment itself. Imagine, for example, that the treatment group in a drug experiment is asked to come to a doctor's office to get an injection each week, whereas the control group does not visit a doctor's office weekly, because they have no need to do so. If, at the end of the study, the treatment group differs from the control group in terms of any outcome measures, the difference could have been caused by the doctor visits experienced uniquely by the treatment group (e.g., because each visit brought the participants into contact with very sick people, and seeing them inspired them to live healthier lives in general), not to the drug itself. Therefore, experimental researchers work hard to minimize the extent of all possible differences between the treatment and control groups, other than the essence of the treatment itself.

Another potential source of confounds are participants' beliefs about the hypothesis being tested and their role in the experimental investigation. Imagine that members of the treatment group are told that the study is investigating whether a new

drug improves cholesterol levels and they that will be receiving the drug, whereas the control group is told about the purpose of the study and is told that they will not be receiving the drug. Simply believing that one is receiving a new and potentially helpful drug might make people happier and more optimistic about their lives, and this general optimism might cause lifestyle changes that themselves improve cholesterol levels. Therefore, experimental researchers believe that it's important for research participants to be as uninformed as possible about which experimental group they have been assigned to, so as to minimize the risk that expectations will cause differences between the groups in health status.

It is interesting to note, however, that the scientists' desire to conduct an experiment with strong internal validity (meaning that he or she can reach a strong conclusion about whether the treatment of interest caused changes in the outcome variable of interest) can limit the external validity of the findings (meaning the degree to which the experiment's results describe what would happen to people experiencing the treatment outside the experiment in the course of everyday life). For example, when conducting an experiment, the researcher might prefer that participants not know whether they are in the treatment or control group, whereas in real, life people taking a particular drug weekly would be well-aware that they are doing so. As a result, the researcher must decide whether he/she is interested in learning about the chemical and biological impact of the drug per se, or whether he/she is interested in learning about that impact coupled with the impact of the knowledge that one is taking the drug.

Applying Experimental Methods to Studying Questionnaires

The use of experiments to study survey questions is based on this same basic philosophy. Such experiments can be done in one of two ways, either by manipulating a question or by manipulating the conditions under which a question is asked. In the first, answers obtained by one version of a question are compared to the answers obtained by a different version of the question. If answers vary, then that variation can be attributed to changes in the question. The second approach involves asking a question of all people but varying the context in which it is asked. Both approaches can yield results that will help to optimize principles of optimal question design.

Experiments with questions can be done employing either a between-subjects design or a within-subjects design. In a between-subjects design, each respondent is randomly assigned to be asked one version of a question or another version of it. Thus, differences between the questions are ascertained by comparing the answers provided by groups of people. This approach can detect aggregate patterns of change, collapsing across people. For example, if some respondents are asked whether they approve or disapprove of a policy, while others are asked whether they favor or oppose it, the proportions of people who offer a favorable response can be compared to ascertain whether one question yields more such responses. If different people's answers are changed in opposite directions (some shifting from favorable to unfavorable, and others shifting from unfavorable to favorable), net change may understate the number of respondents' answers that would have been different had they been asked the other question. With this design, it is also not possible to determine whether any one

respondent's answer would have been different had he or she been asked the different question. However, if a researcher has hypotheses about particular subgroups of respondents being affected differently by a question manipulation, tests can be conducted by comparing responses to the different questions within sub-groups of respondents.

In a within-subjects design, the same respondents are asked more than one version of the same question. In theory, this allows a researcher to identify people whose answers would or would not be different depending on which question they are asked. However, in order for this approach to be informative about how a respondent would answer a single question in isolation (since researchers usually wish to identify the single optimal version of a question to ask), carry-over effects must be eliminated. One type of carry-over effect would involve respondent fatigue: if respondents are asked one version of a question and are then asked a different but very similar version of it, they may doubt the competence or efficiency of the researcher, who might appear to be wasting their time by making the same inquiry twice. Such doubt could undermine respondent motivation to provide accurate answers to the survey questions, thereby compromising data quality in the remainder of the questionnaire.

A second type of carry-over effect would occur if respondents not only remember that they were asked a previous, highly similar question, but also remember the answer they gave to that question. One of the core principles validated by decades of research in psychology is the notion of "commitment and consistency" (Cialdini, 2000; Festinger, 1957). Once people express a particular opinion or describe themselves in a particular way, they have a tendency to want to maintain an image consistent with that initial

statement. Therefore, respondents' answers to a later, similar question may be derived from their recollection of their answer to the earlier question.

A third type of potential carry-over effect derives from the rules of conversation (Grice, 1975). Everyday conversation is governed by a set of principles that people learn implicitly, that speakers follow, and that listeners assume that speakers follow. One such rule is that speakers ask or say only what is necessary. Therefore, a listener can normally presume that anything a speaker says is said because the speaker thinks it is necessary for the conversation to be effective. Asking the same question of a person twice would most likely violate this rule. For example, if I were to ask a person, "How are you?" and she answered "Fine," it would be very odd indeed for my next question to her to be, "How are you?", because she can presume that I already know the answer. If I were to ask the same question again, she might wonder whether I didn't hear her initial answer, or she might think that I'm asking again because her initial answer was an inadequate answer in my opinion, so I seek more elaboration. But if an interviewer were to ask a simple, dichotomous yes/no question in a face-to-face interview, obviously hear and record the respondent's affirmative answer, and then ask the same question again, this would clearly be an unexpected and puzzling violation of the rules of conversation. As a result, if a questionnaire were to ask a question once and then ask a subtly different version of it later, even if the researcher's intent was to measure the same judgment in a slightly different way, the respondent might be motivated to find a way to interpret the second question so that its meaning is as different as possible from the initial question's. That way, it would make sense for the questioner to ask the second question.

Consider, for example, a researcher who wants to compare answers to two questions measuring life satisfaction: “How satisfied are you with your life?” and “How happy are you with your life?” If each respondent is asked only one of these questions, all respondents might interpret both questions as having essentially the same meaning. But if a respondent were initially asked the first of these questions and was then asked the second, he or she might try to find a way to interpret the second question so that it has different meaning than the first. Of course, the difference between the two questions is “satisfied” vs. “happy”. So a respondent might choose to answer the second question not by reporting overall satisfaction with his or her life but instead to report how often he or she feels happy or how happy he or she typically feels, which would be describing the frequency or intensity of a mood state, rather than the degree to which life meets some standard of acceptability, regardless of emotional state. Therefore, the two questions might acquire very different answers from the respondent, even though they were intended to measure the same construct.

If a respondent were to be asked both questions in the reverse order, then the interpretation process would most likely unfold in a very different way (because the respondent would most likely interpret the first, happiness question as tapping overall life satisfaction, then leaving the respondent puzzled about how to reinterpret the second, satisfaction question). None of this is desirable when conducting a within-subjects experiment to compare two questions. Therefore, if a researcher were to implement such an experimental design, he or she should assure that there is sufficient time passage (and perhaps distraction) between the administrations of the two similar questions so as to

minimize the likelihood of contamination of answers to the second by answers to the first.

Unfortunately, it is difficult to know on theoretical grounds or based on practical evidence how long the necessary minimum time interval must be. Van Meurs and Saris (1995) reported evidence suggesting that 20 minutes is sufficiently long. If this is correct for most questions, and if a researcher's questionnaire is not naturally at least 20 minutes long, then achieving the passage of time of 20 minutes would require lengthening an interview or would require making a second contact with the respondent on another, later occasion. Either way, this would substantially increase respondent burden and researcher challenge and thereby increase the cost of implementing a within-subjects experiment. Consequently, the seeming efficiency that this design brings is likely to come at significant practical costs.

However, it seems unlikely that the needed time interval to assure forgetting of a prior question is uniformly 20 minutes. Most likely, this time interval varies as a function of the particular topics and forms of the questions and intervening events. For example, if during a long time interval filled with questions about a person's personal finances, I were to insert two questions, one early and one late, about whether he or she had ever run over a cat or dog with a car by accident, the respondent seems likely to remember this unusual and emotionally provocative question long after 20 minutes have passed. Researchers could address the risk of this latter type of carry-over by, at the time that respondents are asked the second question, also asking whether they remember being asked a similar question earlier, and if so, how that question was worded and how they

answered it. But doing so would then require asking more questions of respondents and would therefore lengthen the interview further.

If researchers were to discover that a substantial number of respondents did remember the prior question and their answers to it, then the value of the experiment would be significantly compromised. The people who remember would need to be dropped from analysis, thereby reducing the representativeness of the sample in a biased way and reducing the effective sample size and statistical power. Or the experiment could be redesigned and rerun, thus throwing away all the data from the first execution, again costly in terms of researcher and respondent time, money, and perhaps other resources.

For all of these reasons, it is not obvious that a within-subjects design has practical or analytic advantages over between-subjects designs. And for me personally, the practical challenges of minimizing carry-over effects are substantial enough so that I strongly prefer between-subjects designs. This preference brings with it not knowing which particular respondents would have answered the two forms of the same question differently. But most question evaluation and comparison can be done without knowing that information, because we usually seek to make generalizations about questions regardless of particular individuals.

Analyzing Experimental Data

When conducting question design experiments, researchers most often focus on two principal criteria: distributions of answers and correlates of answers. For example, one might wonder whether the order of presentation of the answer choices influences

answers to a closed-ended question. To do so, half of the respondents in an experiment could be given the choices in one order, and the other half could be given the choices in the reverse order, and the distributions of answers to the two versions could be compared. No difference in distributions would suggest that there had been no impact of order. The less impact that order has on responses, the more valid the question might appear to be, because if a person's answer to a question is an accurate self-description, then that answer should be the same regardless of the order in which the choices are offered.

A researcher might also wish to assess the validity of measurements obtained by different versions of a question by observing their correlations with other variables. For example, imagine that a researcher were interested in identifying the most valid way to measure life satisfaction and that solid theory and evidence indicate that people who earn more money are happier, on average. A researcher could then measure earnings for a sample of people and randomly assign each of them to get one of two different versions of a question measuring life satisfaction. If answers to the two questions correlate equivalently with income, then they would appear to be equally valid according to this test. But if answers to one version correlate more strongly with income than do answers to the other version, that would suggest the former might be more effective at accurately assessing life satisfaction, because answers contain less random and/or systematic measurement error.

Yet another analytic option is to examine test-retest reliability of items through experiments. To do so, each respondent can be asked the same question two or more times, separated by a suitably long time interval of days, weeks, or months. Then, an

analyst can estimate the consistency of answers to the question as an indicator of reliability of the measurements. If different respondents are randomly assigned to answer different versions of the question, then the reliabilities of the various items can be compared to one another to identify the superior item in terms of measurement quality. Refinement of the analytic approach can be enhanced if each respondent answers multiple questions tapping the same construct on each occasion, so that latent variable covariance structure modeling can be conducted to produce an estimate of the reliability of the items while controlling for any change that might have occurred in the construct between the measurement occasions. This can be accomplished even if only a single question is asked on at least three occasions (see, e.g., Krosnick, 1988).

Statistical Options and Issues

When analyzing data from question design experiments, a number of analytic methods can be used. First, when comparing the distributions of answers, it is common to create a contingency table cross-tabulating question version with answers. For example, in an experiment that varied the order in which response choices were offered, each row can correspond to an answer choice, the columns can distinguish the groups of respondents asked the question with the different orders, and the cell entries can be the percent of people in each experimental group (i.e., column) who gave each response, with each column totaling 100%. The statistical significance of the differences between distributions can be assessed with a χ^2 statistic. If a question offers interval-level response options (e.g., "On how many days during the last week did you eat bread?"), a researcher could compute the mean response to two versions of the question and compute

a t-test to assess the statistical significance of the difference between the two means.

When computing such statistical tests, a researcher must grapple with some potential complexities due to the design of the study. One issue is whether to weight the data or not. Two types of weighting can be done. One is to reflect unequal probabilities of selection to participate in the survey. Imagine, for example, a study in which interviewers visited the homes of respondents and randomly selected one adult resident to interview in each household. Thus, a resident's chances of being selected into the survey are inversely proportional to the number of adults living in the household. In order to properly project a survey's results to the population, it is necessary to statistically adjust for such intended inequalities of probability of selection through weighting. This type of weighting follows unambiguously from the design of a sample and is therefore relatively straightforward to implement.

A second type of weighting is post-stratification. When a survey sample's demographics differ notably from the population of interest in terms of known distributions of benchmark variables, analyses can be done while increasing the weight assigned to respondents from under-represented groups and reducing the weight assigned to respondents from over-represented groups. This sort of weighting is much more an art than an exact, formulaic science. Sampling statisticians do not agree on a single, optimal approach to computing such weights.

Gelman (2007) argued that rather than doing post-stratification weighting using demographics, researchers should statistically control for all such demographics when estimating the parameters of regression equations. This advice might seem relevant to

the analysis of question design experiments, because this can be done with regression. For example, if a researcher wants to assess whether one version of a life satisfaction question yields more favorable answers than another, he or she could randomly assign respondents to be asked one of those two versions and regress answers (either coded continuously and analyzed with OLS regression or coded categorically and analyzed with multinomial regression) on a dummy variable differentiating people who were asked the two different versions of the question plus demographic control variables that could instead have been used to construct post-stratification weights.

In the context of an experiment, controlling for demographics is very unlikely to alter the parameter estimate for the effect of the question design manipulation, because random assignment will mean that the demographic controls will be essentially uncorrelated with the question design manipulation. But controlling for demographics is likely to improve the ability of the analysis to detect the statistical significance of the manipulation's effect if the demographics explain any of the variance in responses. This is likely because controlling for those demographics will reduce the error variance used in tests of statistical significance.

However, controlling for demographics in such an analysis to solving unrepresentativeness in the sample seems unlikely to yield the same outcome as post-stratification would unless the impact of the question manipulation is uniform across the sample. If instead, the size of the effect of the manipulation is moderated by a variable that would have been used in weighting, then additional computation must be done. By failing to take into account the deviations between the sample and the population in terms

of the distribution of the moderating variable, people at various levels of that variable will not be represented in the analysis in their proper proportions.

Imagine a case in which a question manipulation has strong impact on answers provided by respondents who did not graduate from high school and has no impact at all on answers from people who did graduate from high school. And imagine that the participants in an experiment vastly under-represent the proportion of people who did not graduate from high school. If the data were analyzed without post-stratification, the few people in the sample without a high school degree would be vastly outnumbered by the others, and the lack of responsiveness to the question manipulation among the latter individuals could prevent a researcher from seeing a statistically significant effect of the question manipulation in the sample as a whole which is due to the responsiveness of people who did not graduate from high school.

Implementing post-stratification to correct the under-representation of people without high school degrees would greatly increase their presence in the sample and might therefore cause the full sample test of the question manipulation to yield a significant effect. Even if a researcher does post-stratification, controlling for all demographics and other variables that were used to generate the weights is likely to be a wise idea, again because if those benchmarks are related to answers to the manipulated question, then controlling for them will remove some systematic variance in answers and increase a researcher's ability to detect a real difference between the question forms as statistically significant. As long as assignment to question version is done truly randomly, controlling for weighting variables in this way is unlikely to cause an effect of

question design to appear to be significant when in fact it is an illusory artifact of the statistical estimation procedure.

Using the approach of not post-stratifying and instead controlling for the variables in terms of which the sample is known to deviate from the population will produce a proper total experimental effect size only if the researcher makes post-estimation adjustments for variation in the experimental effect size across subgroups of the sample. The effect size must be computed within groups of people differing in their values on any moderating variable (e.g., at different levels of education), and the estimated effect sizes can be combined in a way that takes into account the proportions of the various groups of respondents in the population, so as to yield an overall effect size for the population. The only way to accomplish this is to check for variation in the effect size across all levels of all variables in terms of which the sample deviates from the population. In this light, post-stratification may be a simpler and effective way to accomplish the goal of producing an effect size estimate for the population, if that is of interest.

Such post-stratification does not come without a cost. Specifically, such weighting weakens statistical power because of uncertainty in the weights themselves. The more variable the weight values are from one another across respondents, the greater the "design effect" for the weighting is. The larger the design effect, the more variance in answers is caused by the researcher's somewhat arbitrary decision about how to construct those weights. If data are analyzed properly with statistical software that recognizes the impact of the weights on statistical confidence, weighting reduces statistical power to detect a real difference between answers to two different versions of a question. The

larger the design effect of a survey's weighting approach, the more risk the researcher takes of failing to detect a real influence of a question manipulation on answers.

Therefore, the decision about whether or not to weight data from a question design experiment must be made based upon a researcher's goal for the experiment, and there are at least two legitimate but importantly different possible such goals. The first goal is to ask whether the question manipulation had a reliable effect on the answers provided by the people who participated in the study. This is a legitimate question and in fact has probably been the default assumption made in the vast majority of experiments conducted across the social and physical sciences. The data from such experiments have routinely been analyzed asking whether the two or more experimental groups differed from one another due simply to chance alone due to the random assignment procedure, or whether the manipulations are likely to have caused observed differences between those particular groups of people.

Another, equally legitimate goal would be to ask what the impact of this experiment's findings would have been if the study had been done with a fully representative sample of a particular population. To ask that question requires specifying that population, of course. And it is not always obvious what population should be used for this purpose. For example, if an experiment were to be conducted in 1970 in the United States and researchers wished to ascertain what its result would have been if all American adults had been interviewed at that time instead, it is reasonable to post-stratify the data to match the American adult population in 1970.

But if a researcher wants to make a broader statement about the difference

between the question versions, he or she might be tempted to think that any conclusions reached from this experiment would apply in 1980 and 1990, not only in the United States but in other countries as well. Therefore, it might be tempting to analyze the data weighting them numerous different ways to reflect many different populations, including hypothetical future populations that do not yet exist. Clearly, there is an endless number of possible populations to which a researcher might weight if he or she wishes to make general statements about the impact of a question manipulation. So simply weighting to match one of these many populations is not likely to be especially intellectually satisfying. Nonetheless, doing so might be of some value to reassure scientists that the result of an experiment with a highly unrepresentative sample can be generalized to a very different population to be studied later.

Another analytic issue to consider with experimental data is clustering in the sample design. National, face-to-face surveys routinely keep costs under control by sampling cases in geographically defined clusters (e.g., primary sampling units). As a result, interview locations are not smoothly spread across the entire nation (which would require a lot of interviewer traveling or a huge number of interviewers) and instead are grouped in clusters. Because of the clustering, all people interviewed who live within a single cluster are likely to be more similar to one another than they are to people in different clusters. Routines for representing such non-independence are offered by various statistical packages, and they should be used to properly adjust the error variance to reflect this known source of covariation.

Assessing Validity

Comparing the accuracy of measurements obtained by two versions of a question via predictive validity can be done by measuring the association of the two sets of answers with a criterion variable. Depending upon whether the variables involved have nominal response options or interval-level response options, a researcher can conduct OLS regression or logistic regression or multinomial logistic regression, in addition to other such techniques. Regression coefficients can be computed to gauge the associations between variables, and coefficients for different question versions can be compared.

In a regression is conducted predicting a criterion with answers to a target question, a dummy variable indicating which version of the target question each respondent was asked, and the product of those two predictors, the product tests the interaction and therefore effectively tests whether the relation of the target question with the criterion was significantly different depending on which version of the target question was asked. The version producing the stronger relation is presumed to have produced the more valid measurements.

The ideal criterion variable for comparing the validities of two versions of a question would be a perfect measure of the construct of interest. For example, if one wanted to measure the amount of exposure a person had to the television news programs during the past week, it would be wonderful to correlate answers to two survey questions with a pure and completely accurate assessment of that television news program exposure. With such a measure, we could estimate the parameters of the following

equations separately using two different questions measuring media exposure:

$$\Gamma_1 = b_1 (T) + s_1 + e_1 \quad (1)$$

$$\Gamma_2 = b_2 (T) + s_2 + e_2 \quad (2)$$

where Γ_1 is answers to one question asking about television news exposure, Γ_2 is answers to the second question assessing television news exposure, T is the true amount of television news exposure each respondent experienced, b_1 is the validity of Γ_1, b_2 is the validity of Γ_1, s_1 and s_2 represent systematic measurement error in answers to each question (such as a tendency for people to under-report exposure using a particular measure, either intentionally or accidentally because of misremembering), and e_1 and e_2 represent random measurement errors in answers to each question. If $b_1 > b_2$ and/or $e_1 < e_2$, that would suggest that the first question is a more valid and/or reliable measure of true exposure than the second question. And if $b_1 < b_2$ and/or $e_1 > e_2$, that would suggest that the first question is a less valid and/or reliable measure of true media exposure than the second question.

Unfortunately, no pure and completely accurate assessment of television news exposure or any other construct of interest in surveys yet exists. For example, to measure media exposure, many different approaches have been explored, including observation, diaries, experience sampling, and more, and a large literature has emerged highlighting advantages and drawbacks of them all (for reviews, see, e.g., Engle and Butz 1981; Kubey and Csikszentmihalyi 1990; Robinson and Godbey 1997, p. 61 - 62; Stipp 1975; Webster and Wakshlag 1985). Each of these methods is subject to unique sources of systematic measurement error (e.g., diaries are often not filled out daily but rather are

filled out for the entire reporting period just before they must be turned in), so none is perfect. The validity of global self-reports can be assessed via correlations with such alternative measures, but recognizing that perfect measurement is impossible.

When no direct measure of the construct of interest is available, researchers can take an alternative approach by using a criterion variable that theory suggests should be correlated with the construct being measured by the target question. This approach was suggested by the American Psychological Association (1954) for gauging the validity of a measure: assessing construct validity, which focuses on the extent to which a target measure is related to measures of other constructs to which theory says it should be related (see also Messick 1989).

The relation of two versions of a target question to such a criterion can be represented this way:

$$Y = b_3 (\Gamma_t) + b_4 (\Phi) + s + e \quad (3)$$

$$Y = b_5 (\Gamma_p) + b_4 (\Phi) + s + e \quad (4)$$

where Y is the criterion measure that is associated with television news exposure (e.g., a quiz assessing the amount of factual knowledge about politics that the respondent possesses), Γ_1 and Γ_2 are the two different measures of television news exposure, b_3 and b_5 are coefficients estimating the associations of Y with Γ_1 and Γ_2, Φ is a vector of other correlates of the criterion that have been measured in the survey, b_4 is a vector of coefficients reflecting the strength of impact of these other causes, s is systematic measurement error in assessments of the criterion, and e is random error in measurements of the criterion. b_3 and b_5 can be estimated in the two separate equations leaving Φ and s

out of the equation, because the impact of other causes and systematic measurement error will be the same in both. Invalidity and random measurement error in the measures of television news exposure will attenuate b_3 and b_5. So if $b_3 > b_5$, that would suggest that the first question is a more valid and/or reliable measure of its construct than the second version. And if $b_3 < b_5$, that would suggest that the first question is a less valid and/or reliable measure of its construct than the second question.

Another analytic approach that can be employed to assess item validity in an experiment involves latent variable covariance structure modeling (e.g., Bollen, 1989). This approach can be implemented by collecting multiple identical measures of a construct and randomly assigning respondents to be asked one of various different versions of another measure of the construct. The various measures can then be treated as indicators of a single latent construct in a covariance structure analysis. Software such as LISREL, M-Plus, Amos, or EQS can be used to produce estimates of the validity and reliability of the different versions of the target measure, and tests can be computed to assess the significance of the difference between these parameters.

Such analysis can be done relatively simply by including just the latent construct of interest in the model and various measures of it. However, it is also possible to include measures of other constructs to which the construct of interest is likely to be related. So, for example, a study comparing the validity of various life satisfaction measures can administer all such measures to respondents and also measure their incomes. Income can be represented in the covariance structure model as a separate construct correlated to some unknown degree with life satisfaction. With this model

structure (2 correlated latent constructs: life satisfaction and income), the statistical estimation procedure has more information (i.e., the correlations of the various target measures of life satisfaction with income) to use in gauging the validity of the two alternative versions of the measure of life satisfaction.

When associations between questions are estimated using experimental data, researchers are often tempted to compute standardized measures of association between items, such as Pearson Product Moment Correlations. But comparing the magnitude of a correlation across different versions of a question can be misleading if the questions differ in the variability of responses to them. Consider, for example, two questions that are equally valid in tapping the construct of interest, but one question yields answers that are more variable that the other. This increase in variance will lead a predictive validity correlation to appear stronger for the latter question. It is therefore preferable to examine unstandardized regression coefficients to estimate the strength of associations between items, with all variables coded to range from 0 (meaning the lowest possible level of the construct) to 1 (meaning the highest possible level). Such coefficients are easy to interpret and are less impacted by differences between experimental conditions in the variance of the items involved and therefore provide a clearer comparison of the validities and/or reliabilities of the items.

Assessing Administration Ease

In addition to measuring the impact of question variations on response distributions, reliability, and validity, researchers can conduct experiments to compare the amount of time that it takes respondents to answer different versions of a question.

Response latency, as psychologists call it, can be measured easily under conditions of computer administration by having the computer record the moment in time when a question appears on the screen and the moment in time when the respondent submits an answer to the question. In telephone or face-to-face interviews, the interviewer can push a key on a computer keyboard to mark the moment in time when he or she finishes reading a question aloud and can push the key again to mark the moment in time when the respondent begins to utter an answer to the question. For practical reasons, questions that people can answer more quickly are generally preferred to those that take longer to answer, because asking questions that can be administered more quickly allow for asking more total questions in a fixed interview time period. But in addition, answering a question more quickly may be an indication that the cognitive tasks of interpreting the question, retrieving information from memory to generate an answer, compiling a summary answer, and reporting it were easier for respondents to do, which would make the question desirable, because it would minimize respondent fatigue.

How Experiments Can Be Valuable for Evaluating Questions

Experiments are valuable for evaluating questions because they offer a quantitative technique for assessing whether question variations cause changes in answers, either in terms of their distributions or their validities. If a researcher is uncertain about which question approach is preferable for measuring a construct, an experiment can be conducted to provide evidence with which to decide. And if a researcher has a hypothesis about a particular way in which a target question might bias or distort measurement of the construct of interest, it is possible to compare that target

questions to other versions of it in an experiment to test the hypothesis.

Next, I review illustrations of how experiments have been conducted to help optimize question design. Specifically, I describe experiments that tested for effects of response choice order, question balancing, acquiescence response bias, branching/labeling of bipolar rating scales, word choice, and social desirability response bias.

Response Order Effects

In a study of response order effects, Schuman and Press (1981) asked a randomly selected half of a telephone survey sample this question:

> "Some people say that we will still have plenty of oil 25 years from now. Others say that at the rate we are using up our oil, it will all be used up in about 15 years. Which of these ideas would you guess is most nearly right?"

The other half of the respondents were asked instead:

> "Some people say that at the rate we are using up our oil, it will all be used up in about 15 years. Others say that we will still have plenty of oil 25 years from now. Which of these ideas would you guess is most nearly right?"

Thus, the order in which the two answer choices were read to respondents was varied between subjects. When this experiment was run in January, 1979, 64% of respondents chose the "plenty" option when it was presented first, whereas 77% of respondents chose it when it was presented last, a highly significant difference ($p<.001$). This is what is called a "recency" effect, because a response option is advantaged when it is presented last. This experimental evidence documented a source of systematic measurement error

in responses to the question.

In another study, Holbrook, Krosnick, Carson, and Mitchell (2000) investigated how the order in which response choices are offered can either conform to or violate respondents' expectations based on conversations conventions and can thereby compromise the accuracy of the answers respondents provide. In their experiments, some respondents were randomly assigned to be asked this question:

> "The federal government is considering raising the import tax on steel that comes into the United States from other countries. Raising the steel tax would protect the steel industry from foreign competition and create more jobs for American steel workers. However, it would also increase the prices Americans pay for products made from steel. If you could vote on this, would you vote for raising the import tax on steel or would you vote against it?"

Others were randomly assigned to answer this question:

> "The federal government is considering raising the import tax on steel that comes into the United States from other countries. Raising the steel tax would protect the steel industry from foreign competition and create more jobs for American steel workers. However, it would also increase the prices Americans pay for products made from steel. If you could vote on this, would you vote against raising the import tax on steel or would you vote for it?"

Consistent with the notions that (1) asking whether one favors or opposes in the more natural way to phrase the question, and (2) asking whether one opposes or favors is counter-normative and distracting, Holbrook et al. (2000) found that people answered the

first question more quickly than they answered the second (3.67 sec. vs. 4.04 sec. on average, $p<.05$) and that the predictive validity of questions using the conventional response option order was higher than that of questions employing the counter-normative response option order.

Question Balance

Shaeffer, Krosnick, Langer, and Merkle (2005) explored whether minimal balancing of a closed-ended question is sufficient for producing unbiased responses, rather than having to implement full balancing. To do so, Shaeffer et al. (2005) randomly assigned respondents to be asked one of two versions of the same question:

Minimal Balance. "As it conducts the war on terrorism, do you think the United States government is or is not doing enough to protect the rights of American citizens?"

Full Balance. "As it conducts the war on terrorism, do you think the United States government is doing enough to protect the rights of American citizens, or do you think the government is not doing enough to protect the rights of American citizens?"

The distributions of answers to the two questions were the same, as were the strengths of the associations between these questions and criteria. Therefore, the authors concluded that the two question forms yielded equivalent data, so the minimal balancing approach (which entails fewer words) appears to be the preferable approach.

Acquiescence Response Bias.

Acquiescence response bias is the tendency to provide affirmative answers to

questions offering agree/disagree, true/false, or yes/no answer choices. One experiment illustrating acquiescence was conducted by Schuman and Presser (1981), who asked respondents whether they agreed or disagreed with one of the following two statements:

> "Individuals are more to blame than social conditions for crime and lawlessness in this country."
>
> "Social conditions are more to blame than individuals for crime and lawlessness in this country."

Of the people given the first statement, 60% agree with it, which might lead one to expect that at least 60% of people given the second statement would disagree with it. But in fact, only 43% of people disagreed with the second statement, and 57% agreed with it. Thus, it appeared that a majority of respondents agreement with both a statement and its opposite. This identified a source of systematic measurement error present in answers to these questions.

Branching/Labeling of Bipolar Rating Scales.

Krosnick and Berent (1993) explored the idea that when respondents are asked to make a rating on a bipolar dimension (with a zero point in the middle), reporting accuracy may be improved by decomposing the reporting task into two sub-tasks: reporting whether the respondent is at the mid-point, on one side of it, or on the other side of it, and then separately reporting how extreme the respondent is on his or her chosen side of the midpoint. To do so in one experiment, some respondents (chosen randomly) were asked:

> "There has been a lot of debate recently about defense spending. Some people

> believe that the U.S. should spend much less money for defense. Suppose these people are at one end of a seven-point scale, at point number 1. Others feel that defense spending should be greatly increased. Suppose these people are at the other end of the scale – at point number 7. And, of course, other people have opinions somewhere in between, at points 2, 3, 4, 5, and 6. Where would you place yourself on this scale?"

Others were instead asked this branched question sequence:

> "There has been a lot of debate recently about defense spending. Do you think the U.S. should spend less money on defense, more money on defense, or continue spending about the same amount on defense? [If less:] "Would you say we should spend a lot less, somewhat less, or a little less?" [If more:] "Would you say we should spend a lot more, somewhat more, or a little more?"

Each respondent answered the question twice, separated by about four weeks. Respondents also answered four other questions in the same format, either not branched or branched. Collapsing across the four questions, 39% of respondents gave the same answer to the same question when asked in the non-branching format, and 64% did so when asked the branching format, a highly significant difference ($p<.00001$). Thus, branching and labeling all response options with word improved test-retest reliability and appears to improve measurement quality.

Question Wording.

Chang and Krosnick (2003) investigated whether question wording affected the validity of measurements of exposure to political news through the media. Respondents

in their study were randomly assigned to be asked one of two versions of an exposure question:

"How many days in the past week did you watch the news on TV?"

"How many days in a typical week did you watch the news on TV?"

Similar questions were asked about newspaper reading as well.

Chang and Krosnick (2003) gauged the validity of these items by estimating associations of answers to them with four measures assessing the amount of knowledge each respondent possessed about politics. This approach was based on the assumption that most people gain most of their political knowledge from exposure to the news media, so stronger relations between media exposure and knowledge volume would be an indication of greater validity of the measure of exposure. Political knowledge volume was measured by (1) asking respondents to provide a summary judgment of how informed they were about politics, and (2) giving respondents quizzes asking them to describe specific recent national and international events as best they could; coders graded the accuracy of the information each respondent had on each issue. Chang and Krosnick (2003) found greater predictive validity for the typical week questions than for the past week questions, suggesting that the former produced more accurate assessments of chronic levels of news media exposure.

Social Desirability Response Bias.

Holbrook and Krosnick (2010) explored the impact of social desirability response bias on reports of past behavior. Their focus was on respondent reports of whether they voted in a recent national election. Many scholars have speculated that such reports are

intentionally distorted by respondent desire to appear to have fulfilled their civic duty, since people might be embarrassed to admit that they did not vote. Thus, some people who did not vote might claim to have done so when asked to admit that aloud to an interviewer. Some studies have suggested that such social desirability pressures may be removed when respondents answer questions on a computer, without having to admit their failings aloud to an interviewer (e.g., Chang & Krosnick, 2010).

To test whether social desirability pressures distort reports of turnout, Holbrook and Krosnick (2010) randomly assigned some respondents in a telephone survey and in internet surveys to answer a direct question asking them whether they voted in a recent election, such as:

> "In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you—did you vote in the Presidential election held on November 7, 2000?"

Other respondents were instead asked to report turnout using the Item Count Technique (ICT). Among these individuals, some (chosen randomly) answered this question:

> "Here is a list of four things that some people have done and some people have not. Please listen to them and then tell me HOW MANY of them you have done. Do not tell me which you have and have not done. Just tell me how many. Here are the four things: Owned a gun; Given money to a charitable organization; Gone to see a movie in a theater; Written a letter to the editor of a newspaper. How many of these things have you done?"

Other individuals were instead asked this version of the question:

> "Here is a list of five things that some people have done and some people have not. Please listen to them and then tell me HOW MANY of them you have done. Do not tell me which you have and have not done. Just tell me how many. Here are the four things: Owned a gun; Given money to a charitable organization; Gone to see a movie in a theater; Written a letter to the editor of a newspaper; Voted in the Presidential election held on November 7, 2000. How many of these things have you done?"

The average answer given to the first question can be subtracted from the average answer to the second question to yield the proportion of people who said they voted in the election. The purported advantage of the ICT is that it allows respondents to provide completely confidential reports. Consistent with this reasoning, Holbrook and Krosnick (2010) found in their telephone survey that 72% of respondents said they had voted when asked the direct question, but the estimated turnout rate according to the ICT measure was 52%, a statistically significant decrease ($p<.05$). In the Internet surveys, the direct self-report questions and ICT measurement yielded equivalent turnout rates, suggesting that social desirability pressures did not distort direct self-reports under these measurement conditions.

How Can Experiments Be Misused or Conducted Incorrectly

Two types of problems have sometimes occurred in experimental studies comparing different questioning approaches. One is the failure of random assignment in between-subjects studies. If respondents are given the opportunity to choose which

version of a question they receive, then the selection may be based on a factor that influences answers. So comparisons of answers to the two question forms may not reveal the impact of the question form per se but may instead be attributable to pre-existing differences between the groups of people who answered the two questions. It is therefore important that random assignment be done.

Doing so might seem easy to do in practice, but it has turned out not to be so easy sometimes. For example, I was involved in the design and conduct of a major national survey in which respondents were supposed to be randomly assigned to be asked one of various different versions of questions. After the data had been collected, we inspected the patterns of question assignments and found them to depart so significantly from what would be expected by chance alone that it led us to doubt the effectiveness of random assignment. The programmer who had been responsible for implementing the random assignment insisted repeatedly that the random assignment had been done properly until finally recognizing that in fact, the assignment had not been truly random. Thus, it is important to be vigilant about the details of the procedure for implementing random assignment.

Even if random assignment is properly implemented, there is a non-zero probability that people in different experimental conditions will differ from one another substantially in ways not due to the manipulations implemented. If those pre-existing differences between the experimental groups are related to the outcome variable of interest, this can cause results to be misleading. Therefore, researchers should routinely check their experimental data to see whether people in different experimental groups are

indeed identical in the aggregate in terms of variables that should not have been affected by a treatment. And if the groups are not identical, it is easy and sensible to statistically control for the impact of the unintentionally confounded variable when estimating the effect of the treatment.

Another problem that appears sometimes in write-ups of experiments is focusing on a single manipulation when it is perfectly confounded with another manipulation that could be responsible for apparent differences between experimental conditions. For example, consider the experiment described above by Berent and Krosnick (1993) in which respondents are randomly assigned to be asked a branching question or a non-branching question measuring the same attitude. In that study, the two tested versions of the question differed not only in terms of branching but also in terms of the verbal and numeric labeling of the scale points. The non-branching version presented most rating scale points with numeric labels only, whereas the branched version of the question did not label any response options with numbers and instead labeled them with words. There is reason to believe that verbal labeling of scale points might improve measurement quality, and other studies reported by Brent and Krosnick (1993) showed just that. They found that part of the data quality improvement attributed to branching in the above example was due to branching, and part was due to verbal labeling of scale points. It would be inappropriate to ignore one of the question format variations and presume that differences between experimental conditions are attributable to the other format variation.

A final mistake that can be made in analyzing the effect of an experimental manipulation is controlling for a variable that was affected by the manipulation of

interest. For example, imagine that a researcher is interested in whether adding an argument to a question changes the answers that people give to it. To do so, some respondents might be asked this question, which is taken from a Time Magazine poll done in June, 2008:

> "There is a type of medical research that involves using special cells, called embryonic stem cells, that might be used in the future to treat or cure many diseases, such as Alzheimer's, Parkinson's, diabetes, and spinal cord injury. It involves using human embryos discarded from fertility clinics that no longer need them. Some people say that using human embryos for research is wrong. Do you favor or oppose using discarded embryos to conduct stem cell research to try to find cures for the diseases I mentioned?"

In order to ascertain whether the penultimate sentence influenced answers, other respondents could be asked the question without that sentence:

> "There is a type of medical research that involves using special cells, called embryonic stem cells, that might be used in the future to treat or cure many diseases, such as Alzheimer's, Parkinson's, diabetes, and spinal cord injury. It involves using human embryos discarded from fertility clinics that no longer need them. Do you favor or oppose using discarded embryos to conduct stem cell research to try to find cures for the diseases I mentioned?"

Imagine that these questions were followed by another question that was asked identically of all respondents:

> “In general, do you think medical research is ever immoral?”

It is easy to imagine that answers to this latter question might be influenced by the presence or absence in the prior question of the sentence, "Some people say that using human embryos for research is wrong." Perhaps people who hear that sentence in the previous question are more likely to say that medical research is sometimes immoral. A researcher might be tempted to control for answers to the question about immorality when assessing the impact of the sentence about using embryos being wrong on answers to the question containing it. This temptation might occur because the researcher thinks it would be advantageous to control for the more general opinion when examining the measure of a more specific opinion. But doing so would be inappropriate, because the control variable is not purely exogenous and is instead influenced by the manipulation of interest.

Coordinating Experimentation With Other Question Evaluation Methods

Although experimentation has yielded hundreds of valuable publications informing optimal questionnaire design on its own, it has the potential to be used in coordination with other question evaluation methods in constructive ways. Specifically, conventional pretesting, behavioral observation, and cognitive pretesting are all evaluation methods intended to identify suboptimalities in the design of questions. When a researcher collects data using such methods and identifies a potential deficiency in a question, he or she often then redesigns the question and pretests it again, to see if the change produced an improvement in the quality of the responses. And most often, this improvement is assessed in terms of whether it was easier to administer the question and/or whether respondents' stated interpretations of the question and/or their articulated

thoughts conform to researchers' hopes or expectations.

Much more rare is to see the results of such pretesting subjected to experimental evaluation. Specifically, an experiment could be conducted in which some respondents are randomly assigned to answer the original, possibly flawed version of the question, and other respondents are instead asked the new, presumably improved version of the question. If the question alteration was indeed an improvement, then we should see differences between the experimental conditions indicating that, such as faster reaction time, greater test-retest reliability of answers, and/or greater predictive validity. This seems like a fruitful direction for future studies to coordinate the use of multiple question evaluation methods. Doing this sort of validation would certainly slow down the process of question design and refinement, but it would be valuable for validating the pretesting techniques themselves.

Methodological Criteria for Including Results of Experiments in Q-Bank

Experiments are likely to produce valuable insights into optimizing question design as long as (1) random assignment is done properly, (2) manipulations are done so that key elements of question design are unconfounded from irrelevant variables, and (3) data are analyzed properly to identify the impact of questions on administration difficulty and/or data quality. Some scholars have shown a preference for studies done of representative samples of populations over studies done with convenience samples, such as groups of college students. I believe that there is scientific value in experimental studies of many different types of respondent groups, even when they are not representative samples of populations. If data are collected from a convenience sample

for one experiment, it is incumbent on the researcher to attempt to replicate the findings of that study with representative samples of populations. If replication occurs and findings are consistent across the various methods, this increases confidence in any conclusions drawn about best practices in question design. Q-Bank can be a place to gather reports of all such experimental studies and compare their results to draw general conclusions about best practices.

Conclusion

Experimentation has always played a central role in all types of scientific investigation, and it has played a central role in research optimizing questionnaire measurement as well. I look forward to many more decades producing hundreds more experiments and just as many valuable insights that will help social science to fulfill its potential by testing interesting and important theories using maximally accurate measurements of the constructs of interest.

References

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51, pt. 2.

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Chang, L., & Krosnick, J. A. (2003). Measuring the frequency of regular behaviors: Comparing the 'typical week' to the 'past week.' Sociological Methodology, 33, 55 - 80.

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. Public Opinion Quarterly, 74, 154 - 167.

Cialdini, R. (2000). Influence: Science and practice. Allyn & Bacon.

Festinger, L. (1962). A theory of cognitive dissonance. Stanford, CA: Stanford University Press.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. Statistical Science, 22, 153 - 164.

Grice, H. P. (1975). Logic and conversation. In P. Cole (Ed.), *Syntax and semantics, 9: Pragmatics* (pp. 113–128). New York: Academic Press.

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly, 74, 37 - 67.

Holbrook, A. L., Krosnick, J. A., Carson, R. T., & Mitchell, R. C. (2000). Violating conversational conventions disrupts cognitive processing of attitude questions. Journal of Experimental Social Psychology, 36, 465 - 494.

Krosnick, J. A. (1988). Attitude importance and attitude change. Journal of Experimental Social Psychology, 24, 240 - 255.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. American Journal of Political Science, 37, 941 - 964.

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational measurement (pp. 13 - 103). New York: Macmillan.

Schaeffer, E. M., Krosnick, J. A., Langer, G. E., & Merkle, D. M. (2005). Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. Public Opinion Quarterly, 69, 417 - 428.

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.

van Meurs, L., & Saris, W. E. (1995). Memory effects in MTMM studies. In Saris, W.E., & A. Münnich (Eds.), Multitrait Multimethod approach to evaluate measurement instruments. Budapest, Eötvös University Press, 89 - 103.