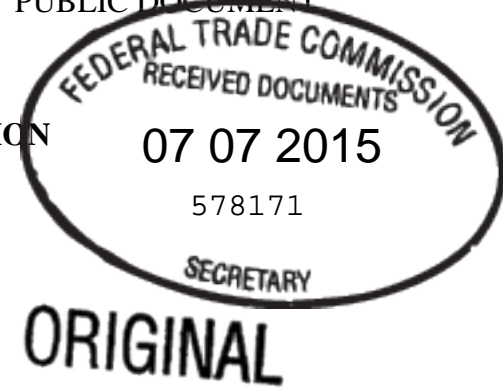


**UNITED STATES OF AMERICA
BEFORE THE FEDERAL TRADE COMMISSION**



COMMISSIONERS: **Edith Ramirez, Chairwoman**
 Julie Brill
 Maureen K. Ohlhausen
 Joshua D. Wright
 Terrell McSweeney

In the Matter of

**ECM BioFilms, Inc.,
a corporation, also d/b/a
Enviroplastics International,**

Respondent.

Docket No. 9358

PUBLIC DOCUMENT

ECM’S RESPONSE TO COMPLAINT COUNSEL’S SUPPLEMENTAL BRIEF

Respondent ECM BioFilms, Inc. (“ECM”) hereby files its Response to Complaint Counsel’s Supplemental Brief (“Br.”).

SUMMARY

Dr. Frederick’s surveys are neither “causal” nor “experimental.” Indeed, they fail to satisfy the accepted requirements for valid and reliable survey research whether “causal” or not. The Synovate survey, relied upon by Complaint Counsel, has previously been rejected by the Commission as unreliable to draw any real world conclusions. RX 348 at 121. The comparison Complaint Counsel offered between the APCO and Frederick surveys is flawed. As Dr. Stewart explains, the two questions drawn by Complaint Counsel from those surveys differ fundamentally; the attempt to compare data sets derived from two different questions yields no sound causal data. Exh. A (Stewart Declaration) at ¶¶18–19. Even a good meta-analysis of badly designed studies will still result in bad statistics. Exh. A at ¶22 n. 23.

Complaint Counsel cite no precedent in which convergent validity theory has been accepted in survey assessment. Complaint Counsel's approach is thus unsupported and, given the flaws in the Synovate, APCO, and Frederick surveys, is illogical and unsound.

A. The Record Contains No Reliable Causal Data

1. Dr. Frederick's Flawed Survey Is Unreliable

Complaint Counsel deem Dr. Frederick's survey "a classic experimental study." Br. at 3–5. It is not. They concede that a "methodologically sound" survey must "draw[] valid samples from the appropriate population, ask[] appropriate questions in ways that minimize bias, and analyze[] results correctly." *Id.* at 3. Dr. Frederick failed to satisfy those criteria. ALJID at 189–202; ECM Answering Br. at 25–34. He failed to ensure that his sample was representative of a defined population. ALJFF ¶¶409–430. He failed to ask appropriate questions in ways that minimize bias. ALJFF ¶¶382–89. He failed to analyze results correctly because, *inter alia*, he flatly rejected one third of all responses. ALJFF ¶¶390–408. His survey "cannot be characterized as . . . valid," ALJFF ¶432, and his results "cannot be relied upon to draw any conclusions." ALJFF ¶434.

Dr. Frederick's survey failed to make the full array of response options available to respondents. ALJFF ¶¶392–93 (explaining Dr. Frederick's "bright-line" rule). Contrariwise, Dr. Stewart's survey, which made all response options available to respondents, revealed that 98% of consumers recognize differences in the amount of time products take to biodegrade, based on the kind of plastic and the environment. Exh. A at ¶6. Dr. Frederick's survey (and the APCO and Synovate surveys) never allowed respondents to explain that biodegradation "depends" on such variables. ALJFF ¶¶392–93, 461–463, 485–86. Dr. Frederick's failure to code non-numeric

responses created a “question constraint,” making his purportedly open-ended questions indistinguishable from closed ones. Exh. A at ¶5.

For open-ended questions to provide causal data, “it should be evident to respondents that they can answer in their own words and what type of response they should provide.” Exh. A at ¶6. But Dr. Frederick generally asked, “[w]hat is your best estimate of the amount of time ...,” begging a specific time response without first discerning whether the respondent believed a specific time response appropriate. Dr. Frederick’s questions thus begot nonsensical answers like “1 nanosecond” and “millennium.” CCX 863. Significantly, Dr. Frederick rejected one third of responses (like “don’t know”) without forewarning respondents that answers that lacked a temporal and numeric unit would be rejected. ALJFF ¶¶392–93. That tactic “provided [Dr. Frederick] with opportunity to interpret and ... ignore responses ... inconsistent with the outcome he was seeking,” resulting in a substantial interpretive bias. Exh. A at ¶7.

By contrast, Dr. Stewart did not restrict answers to specific criteria but allowed respondents to answer in their own words. So, in Question 1 of his survey (i.e., “when you hear the word biodegradable, what does that mean to you?”), he adduced a complete picture of consumer perception, not one constricted through biased questioning. Dr. Stewart’s Question 1 is the only true open-ended observational question before the Commission, the only one that enables us to comprehend what consumers think “biodegradable” means. In response to that, only three percent (3%) equated the term “biodegradable” with a rate of, or time for, biodegradation. RX 605 at 7.

Dr. Frederick’s survey is, at best, a “pseudo experiment.” Exh. A at ¶8. Pseudo-experiments “possess some characteristics” of a true experiment “but suffer design problems which prevent causal inferences.” *Id.* at ¶9.

“Instrumentation” flaws pervade Dr. Frederick’s work, precluding his survey from achieving “causal” status. Instrumentation refers to “problems of measurement that invalidate causal inferences.” *Id.* The first instrumentation flaw is his use of constrained questions. Dr. Frederick’s survey (and the APCO and Synovate surveys) asked respondents how long a product will take to biodegrade, without allowing respondents to answer “it depends” or provide other qualifications. *Id.* at ¶12; ALJFF ¶¶392–93, 461–463, 485–86.

Dr. Frederick’s survey never asked respondents what “biodegradable” meant or screened them for knowledge concerning that essential term. Exh. A at ¶11; ALJFF ¶415. Indeed, he asked just one question per respondent and *no screening questions at all*. “It is well-established that many survey respondents will answer questions even when they do not have any basis for doing so.” Exh. A at ¶11. Because Dr. Frederick asked just one question, we cannot know how each respondent (whether within the so-called “control” or “test” group) defined “biodegradable,” or whether any respondent to Dr. Frederick’s survey even understood the term.

Void of any screening questions, Dr. Frederick’s survey necessarily accepts valid responses from people who have no clue what “biodegradation” means, and data of that kind lack requisite reliability, disqualifying Frederick’s survey for any use, let alone for “causal” or “experimental” use.¹ *Id.*

Indeed, Dr. Frederick’s analysis is invalid because he only included responses consistent with the answers he desired. *Id.* at ¶12. His questions and response options (or lack of response options) likely caused the responses. The stimuli that Dr. Frederick presented, such as a bag labeled “biodegradable,” probably did not affect responses as much as the structure of his

¹ “Over specificity [in Dr. Frederick’s survey] is another example of instrumentation bias: ‘A survey question is overly specific when it asks for an actual or precise response that the respondent is unlikely to know or unable to express.’” Exh. A at ¶11.

questions. Consequently, his causal inferences are highly suspect. *Id.* (citing Patzer) (explaining that “[i]f extraneous variables cause or even partially influence the data in an experiment, subsequent conclusions and actions will likely be erroneous... Conclusions about cause-and-effect relationships, without careful attention to extraneous variables, are generally suspect.”).

Finally, Dr. Frederick erroneously maintains that two materially different questions may be used to generate single “control” and “test” causal data. He offers no support for that approach. *See, e.g.*, Frederick Declaration at ¶¶6(c), 18. His purported control questions, Questions 3H and 3I, asked for a “best estimate of the amount of time it would take for this [container/plastic package] to biodegrade.” CCX 860 at 31–32. His purported test questions, Questions 3J and 3K, asked for a “best estimate of the amount of time it would take for this [container/plastic package] which bears the symbol ECM biodegradable to biodegrade.” CCX 860 at 32–33 (emphasis in original).² Researchers cannot derive single causal data from two materially different questions, because changing the wording “is essentially changing instruments.” Exh. A at ¶15. If a researcher uses different instruments (*i.e.*, questions), comparisons between the instruments are not valid because the researcher cannot know what caused changes in responses. *Id.* A researcher must, of course, know what caused a change in responses to obtain causal data. *Id.* Here the question structure (*i.e.*, adding “which bears the symbol ECM biodegradable”), not the ECM logo or the image, was likely the causal trigger for respondent answers. *Id.* That is, the marketed claim appearing on the bag did not cause the belief, the manipulation of text in the question did. Adding “which bears the symbol ECM

² Those so-called “test” questions repeated the word “biodegradable” three times (once in the picture and twice in the question) in an unnatural way not shown to consumers in the market. The questions thus highlighted the “biodegradable” claim, whereas the so-called “control” question did not.

biodegradable” to the questions rendered them demonstrably leading. *See* ECM Supplemental Br. at 5.³

2. The Synovate Survey Data Is Unreliable

The Commission explained that “[r]eliable real world conclusions cannot be drawn from the Synovate study.” RX 348 at 121. Contradicting the Commission, Complaint Counsel now argue that Synovate’s “results demonstrate that the presence of a biodegradable claim on a plastic item causes a significant number of consumers to believe it will break down in five years or less.”⁴ Br. at 6. No “causal” inferences can be drawn from Synovate questions 8 and 19 because, as even Dr. Frederick admitted, “Synovate #8 is not a perfect control.” Br. at 6–7 n. 4. Synovate Question 8 did not use the same wording as Question 1; did not use the same question stem as Question 19; and offered different response options than Question 19. *Id.*

3. Even A Good Meta-Analysis of Badly Designed Studies Still Results in Bad Statistics

Complaint Counsel argue that Question 3L of Frederick’s survey “can act as a control for” APCO question 4. Br. at 7–8 (citing ¶18 of Frederick Decl.). Dr. Frederick never stated that Question 3L in his survey could serve as a valid control for question 4 in the APCO survey. Rather, he said that the comparison “is not a true experiment.” Frederick Decl. at ¶18. Question 3L in Dr. Frederick’s survey asked “how long will it take to decompose,” whereas Question 4 in the APCO survey asked “what should be the maximum amount of time that it should take for

³ Surveys are causal only where the “experiment is representative of what actually transpires in the marketplace.” ECM Supplemental Br. at 4. By adding “which bears the symbol ECM biodegradable” to his question stems, Dr. Frederick’s questions deviated materially from market reality.

⁴ Whether a biodegradable claim implies to consumers that the plastic will break down in “five years or less” is excludable new argument first made on appeal. Below Complaint Counsel consistently maintained that “biodegradable” implied complete decomposition within “one year” (not more) after customary disposal. ECM Answering Br. at 13–14

that package to decompose?” Br. at 7. As Dr. Stewart explained, citing Bernard, “changing the wording of questions in a survey is essentially changing instruments,” which is an unacceptable survey practice. Exh. A at ¶ 18; *supra* at 5–6. Question 3L in Frederick’s survey is not a control for Question 4 in the APCO survey.

Dr. Frederick provides no support for Paragraph 6(c) of his declaration wherein he suggests that meta-analyses can be used to draw valid inferences from *flawed* surveys. The Frederick survey was invalid. ALJID at 189–202; ECM Answering Br. at 25–34. All agreed that the APCO survey was flawed. ALJFF ¶¶ 456–57, 466–79. A good meta-analysis of badly designed studies will still result in bad statistics, Exh. A at ¶ 22 n. 23 (citing Slavin), and a “meta-analysis” cannot be used to compare Question 4 of APCO and Question 3L of Dr. Frederick’s survey.

B. The Theory of Convergent Validity Does Not Apply to Flawed Studies

Complaint Counsel’s argument that “convergent validity can . . . validate the results of different studies, using different methodologies, conducted at different times by different researchers” is unsupported in the survey literature, illogical, and unprecedented. Br. at 11. Complaint Counsel rely on two cases as support for that proposition, neither one of which applies “convergent validity” theory to survey evidence: *K.S. ex rel. P.S. v. Fremont Unified Sch. Dist.*, 679 F. Supp. 2d 1046 (N.D. Cal. 2009) and *U.S. v. Montgomery*, 2014 WL 1516147 (W.D. Tenn. Jan. 28, 2014)). The expert in *K.S.* reviewed evaluations of a specific student to conclude that he was “‘severely’ cognitively impaired.” 679 F. Supp. 2d at 1052. The court held that the method employed—“reviewing plaintiff’s records and forming a conclusion”—was valid, even though the expert may have “wrongly labeled her analytical technique” as

“convergent validity.” *Id.* at 1060. Similarly in *Montgomery*, surveys were not in issue; IQ tests on a single individual were, and in that context “the principle of convergent validity” was employed to conclude that the person met “the intellectual functioning prong of intellectual disability.” 2014 WL 1516147 at *29. The cases are inapposite. There is no precedential support for the proposition that different surveys, each one flawed, may nevertheless be accepted as valid bases for “convergent validity” analysis.

The ALJ correctly concluded that “[i]t defies logic to contend that three flawed surveys can somehow rehabilitate one another ...” ALJID at 211. Dr. Stewart explained that “the Frederick survey is useless” and “the APCO and Synovate studies are of limited value.” ALJFF ¶22.

Complaint Counsel also argue, citing paragraph 24 in Frederick’s declaration, that all four studies “yield qualitatively similar result[s].” Br. at 12. They do not. At no point has Complaint Counsel or Dr. Frederick explained how the different results in these studies are in fact similar, other than to make that conclusory assertion. Br. at 12; Frederick Decl. at ¶24. Complaint Counsel have not answered the Commission’s essential question, which was to “calculate the degree of convergence...” among the studies. The ALJ considered that same point and concluded that “the evidence does not show that results of the three surveys are similar . . . such that convergent validity theory would even be applicable.” ALJID at 211. The studies are divergent, not convergent (except in the sense that all are flawed). A conclusion of convergence among flawed studies is illogical and invalid.

Respectfully submitted,

s/ Eric Awerbuch
Jonathan W. Emord
Peter A. Arhangelsky
Bethany R. Kennedy
Eric J. Awerbuch
EMORD & ASSOCIATES, P.C.
11808 Wolf Run Lane
Clifton, VA 20124
Telephone: 202-466-6937
Facsimile: 202-466-6938

DATED: July 7, 2015

CERTIFICATE OF SERVICE

I hereby certify that on July 7, 2015, and pursuant to 16 CFR § 4.4(e), I caused a true and correct copy of the foregoing to be served as follows:

One electronic copy and one copy through the FTC's e-filing system to the **Office of the Secretary:**

Donald Clark, Secretary
Federal Trade Commission
600 Pennsylvania Ave., NW, Room H-159
Washington, DC 20580

One electronic copy to the **Office of the Administrative Law Judge:**

The Honorable D. Michael Chappell
Administrative Law Judge
600 Pennsylvania Ave., NW
Suite 110
Washington, DC, 20580

One electronic copy to **Counsel for the Federal Trade Commission:**

Katherine Johnson
Complaint Counsel
Federal Trade Commission
kjohnson@ftc.gov

Elisa Jillson
Complaint Counsel
Federal Trade Commission
ejillson@ftc.gov

Date: July 7, 2015

/s/ Eric Awerbuch
Attorney

EXHIBIT A

**UNITED STATES OF AMERICA
BEFORE THE FEDERAL TRADE COMMISSION**

COMMISSIONERS: **Edith Ramirez, Chairwoman**
 Julie Brill
 Maureen K. Ohlhausen
 Joshua D. Wright
 Terrell McSweeney

In the Matter of

**ECM BioFilms, Inc.,
a corporation, also d/b/a
Enviroplastics International,**

Respondent.

Docket No. 9358

PUBLIC DOCUMENT

**DECLARATION OF DR. DAVID W. STEWART IN RESPONSE TO COMPLAINT
COUNSEL'S AMENDED SUPPLEMENTAL BRIEF**

1. I am David W. Stewart. I have previously provided an expert report, deposition testimony, and trial testimony in the matter of the Federal Trade Commission v. ECM BioFilms, Inc. My prior expert report included statements of my qualifications and described and provided the results of a survey of consumers' understanding of the meaning of the term biodegradable. My earlier report also included a copy of my *curriculum vitae*.

2. I have been asked by counsel for ECM Biofilms to provide responses to the Amended Supplemental Brief filed by Complaint Counsel, which includes my assessment of Dr. Frederick's declaration. Below I present my responses.

3. In Section A(2)(A) of their supplemental brief, Complaint Counsel argue that one third of respondents to Dr. Frederick's survey understood that an unqualified biodegradable claim meant breakdown within five years. That argument, which Complaint Counsel supported

primarily through Dr. Frederick’s flawed and unreliable study, is undermined and contradicted by more reliable data.

4. Ninety-eight percent of the respondents in my survey indicated that they understood that there to be differences in the amount of time it takes for different types of products to biodegrade, decompose or decay, dependent on different kinds of plastics and the environment. (Q4a). That variability extends to plastic products, as there are many types and styles of plastic consumer goods. Respondents in the APCO, Synovate, and Frederick surveys were not provided with a response option that allowed them to offer a qualified response. Indeed, in the Frederick survey, if respondents did offer such a response, it was ignored. That choice of design and exclusion preference violate basic criteria essential for a valid survey. A necessary requirement of valid surveys is that all response alternatives be available to a respondent.¹ This is referred to as being “exhaustive” with respect to response alternatives.

Owens defines “exhaustive” as:

Exhaustive is defined as a property or attribute of survey questions in which all possible responses are captured by the response options made available, either explicitly or implicitly, to a respondent. Good survey questions elicit responses that are both valid and reliable measures of the construct under study. Not only do the questions need to be clear, but the response options must also provide the respondent with clear and complete choices about where to place his or her answer. Closed-ended or forced choice questions are often used to ensure that respondents understand what a question is asking of them. In order for these question types to be useful, the response categories must be mutually exclusive and exhaustive.²

5. The absence of an exhaustive set of potential responses, referred to as “question

¹ Robert B. Settle and Pamela L. Alreck (2004), *The Survey Research Handbook*, (New York: Irwin), p. 110; See Attachment 1. Gilbert A. Churchill, Jr. and Dawn Iacobucci (2005), *Marketing Research*, Ninth Edition (Mason, OH: South-Western), p. 245. See Attachment 2.

² Linda Owens (2008), “Exhaustive,” *Encyclopedia of Survey Research Methods*, (Thousand Oaks, CA: Sage), pp. 248–249. See Attachment 3.

constraint,”³ is not unique to closed-ended questions. It also exists in open-ended questions.⁴ For example, in one study, an open-ended question asked respondents to name the most important events of the last century.⁵ That question produced few mentions of “the invention of the computer.”⁶ However, when the invention of the computer was included in a list of alternative answers to a closed-ended question, it was the most frequent response and was offered as an answer more often than even World War II.⁷ Further investigation in this study revealed that the invention of the computer was indeed considered most important and that the structure of the open-ended question had suggested that respondents name political events rather than other types of events, such as technological innovation.⁸ This example is analogous to the present case in which the structure of Dr. Frederick’s questions constrained answers that became very clear when questions were asked in a different manner, such as in the survey I conducted. Dr. Frederick’s questions were designed to elicit responses about a temporal limitation for biodegradation without first considering whether that temporal limitation was relevant or supported in the study population.

6. Given that 98% of consumers understand that there is no uniform time for products to biodegrade, decompose or decay, the absence of such a response in Dr. Frederick’s survey is a fatal flaw. Indeed, Dr. Frederick made no effort to be clear about what types of responses were acceptable, and was not clear about how the respondents could respond. That is

³ Howard Schuman and Stanley Presser (1996), *Questions & Answers in Attitude Surveys*, (Thousand Oaks, CA: Sage), pp. 299–301. See Attachment 4.

⁴ Norbert Schwarz, Robert M. Groves, and Howard Schuman (1998), “Survey Methods,” in *Handbook of Social Psychology*, Volume 1, Fourth Edition, (New York: McGraw-Hill), p. 160. See Attachment 5.

⁵ Howard Schuman and J. Scott (1987), “Problems in the Use of Survey Questions to Measure Public Opinion,” *Science*, pp. 957–59. See Attachment 6.

⁶ *Id.*

⁷ *Id.*

⁸ *Id.*

contrary to acceptable survey research practice. For example, the highly regarded Pew Center for Survey Research states:

It is important to ask questions that are clear and specific and that each respondent will be able to answer. If a question is open-ended, it should be evident to respondents that they can answer in their own words and what type of response they should provide (an issue or problem, a month, number of days, etc.).⁹

7. Dr. Frederick's survey did not do so, which provided him with an opportunity to interpret and even ignore responses that were inconsistent with the outcome he sought. It also created a substantial bias in his data:

If the criteria by which respondents must judge some issue or respond to some question aren't completely obvious, the criteria must be stated in the question. If an item might be judged by multiple standards and the criteria aren't explicitly stated, some respondents will use one set of criteria and others will use another.¹⁰

In contrast, Question 1 of my survey did not force respondents to answer any question with any specific criteria. That question asked "when you hear the word biodegradable, what does that mean to you?" RX 602. In response, only three percent (3%) equated the term "biodegradable" with a rate of, or time for, biodegradation. RX 605 at 7. As I did not restrict respondents' range of responses, I did not need to state on which criteria the responses were being judged.

8. In Section A(2) of their supplemental brief, Complaint Counsel argue that Dr. Frederick's survey was an experimental survey with appropriate test and control groups. That is plainly contrary to accepted survey research. Dr. Frederick's survey suffered from design problems preventing it from producing causal data.

9. Dr. Frederick's survey was not a classic experiment. Rather, it was what is

⁹ <http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/> (last visited July 2, 2015). See Attachment 7.

¹⁰ See Settle & Alreck, *supra* n. 1 at p. 95. See Attachment 8.

commonly referred to as a *pseudo-experiment*.¹¹ Such “experiments” possess some of the characteristics of a true experiment but suffer design problems which prevent causal inferences.¹² A critical flaw of the Frederick survey and his other analyses is referred to as “instrumentation.”¹³ That limitation on the validity of a research design most often occurs because of changes in measures over time (it refers to any problem of measurement that invalidates causal inferences). There were four fatal errors in Professor Frederick’s analysis that are associated with instrumentation. The first error was the use of constrained questions in his survey, which is discussed above. For a survey or an experiment to be a valid measure of outcomes (the dependent variable) it must enable respondents to provide the full range of potential answers. If a green and red ball are presented to respondents and they are asked whether the ball is green or red, and cannot answer both, the results would clearly be invalid. That was the case with the Frederick survey. It is also not idle speculation that there were many other responses available for Dr. Frederick’s questions since my survey demonstrated that 98% of respondents (correctly) believe that the amount of time for decay depends on a variety of factors. Note that this same problem is associated with the APCO and Synovate surveys.

10. In addition, as noted above, asking questions with an unstated criterion will result in respondents answering quite different questions based on their own idiosyncratic interpretation of the criterion.¹⁴ There is ample evidence of this type of respondent behavior in Dr. Frederick’s survey.

¹¹ See *id.* at pp. 407–09. See Attachment 9.

¹² My earlier affidavit (and expert report and testimony) in the present matter reviewed such design problems in the Frederick survey(s).

¹³ William Shadish, Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, (New York: Houghton Mifflin Company), pp. 60–61. See Attachment 10.

¹⁴ See Settle & Alreck, *supra* n. 1 at p. 95. See Attachment 8.

11. The second limitation of Dr. Frederick's survey was that respondents were not screened for knowledge about biodegradability. It is well-established that many survey respondents will answer questions even when they do not have any basis for doing so.¹⁵ That is why standard practice requires screening of potential respondents based on knowledge of the subject.¹⁶ Respondents in Dr. Frederick's survey were asked questions, and expected to provide a high degree of specificity, without first determining whether respondents had any basis for answering such questions. Respondents in Dr. Frederick's survey were not even asked if they had general familiarity with the term biodegradable, as respondents in my survey were. One might also ask laypersons how quickly Amoxicillin will cure a cold. Many respondents are likely to offer a response but that response will not be based on any understanding that Amoxicillin is an antibiotic rather than an anti-viral medication or of its physiological effects which affect time for cold relief. That is another example of a question suggesting a type of answer. The fact that answers are given and can be tabulated does not make them relevant to the question of the effect(s) of Amoxicillin and is certainly of no use for informing the claims that may or may not be made for Amoxicillin because respondents lacking any knowledge of Amoxicillin (or biodegradation) would simply be guessing. Even if Dr. Frederick's survey questions had not been constrained and leading, they asked for a level of specificity that most respondents were unlikely to be able to address. Over specificity is another example of instrumentation bias: "A survey question is overly specific when it asks for an actual or precise

¹⁵ Patricia Labaw (1980), *Advanced Questionnaire Design*, (Cambridge, MA: Bat Books), pp. 88–92. *See* Attachment 11.

¹⁶ Seymour Sudman and Norman M. Bradburn (1982), *Asking Questions*, (San Francisco: Jossey-Bass), pp. 88–118. *See* Attachment 12; Churchill, Jr. & Iacobucci, *supra* n. 2 at pp. 239–240. *See* Attachment 13.

response that the respondent is unlikely to know or unable to express.”¹⁷ Dr. Frederick asked no screening questions, and asked just one question of each respondent. Thus, we cannot know whether his respondents had any understanding or exposure to the word “biodegradable,” whether they had a knowledge base for answering such a specific question, or how those respondents would define that term.

12. A third measurement or instrumentation problem with Dr. Frederick’s analysis was that he only considered responses consistent with the answers he was seeking in his own survey and focused on a narrow set of answers in the APCO and Synovate studies. For instance, he only considered responses that included both a temporal unit and numerical specification when analyzing his data. Similarly, the response options in both APCO question 4 and Synovate Question 19 were limited, and did not allow respondents to answer that time for biodegradation “depends.” A very strong alternative explanation for the results of these surveys is that the survey *questions* are the causal agent of responses, not the stimuli that Dr. Frederick presented or actual consumer beliefs in the cases of the APCO and Synovate studies. In other words, the structure of the survey questions are responsible for generating a temporal concept of biodegradation that would otherwise not have existed. Dr. Frederick has no way to refute this explanation, and for this reason any causal inference is suspect. The structure of Dr. Frederick’s questions was thus an extraneous but potentially real alternative explanation for Dr. Frederick’s results, as were those obtained in the APCO and Synovate studies, especially in context with the results of the survey I conducted. But,

[i]f extraneous variables cause or even partially influence the data in an experiment, subsequent conclusions and actions will likely be erroneous... An extraneous variable poses a threat to experiments. The threat is that researchers are interested in the effect caused by the controlled and

¹⁷ See Settle & Alreck, *supra* n. 1 at pp. 97–98. See Attachment 14.

manipulated variables under study, but other variables can confuse or confound the data... Extraneous variables in marketing research always pose threats to the accuracy of conclusions based on experiments... Conclusions about cause-and effect relationships, without careful attention to extraneous variables, are generally suspect.¹⁸

13. Indeed, the only available responses to the questions in the APCO and Synovate surveys, other than “don’t know,” were scientifically erroneous based on evidence regarding biodegradability.¹⁹ Similarly, the only acceptable responses in Dr. Frederick’s survey had to be expressed in terms of a specific time interval, which is also inconsistent with scientific evidence. Thus, the most likely explanation for the findings of Dr. Frederick is the structure of the questions he used or examined and NOT any claim at issue in the present matter.

14. Finally, Dr. Frederick compares quite different questions in his test versus control conditions. In his control conditions his survey asked for a best estimate of how long it will take for the container/package to break down. In the test conditions he asks for a best estimate of how long it will take for this container/package which bears the symbol ECM biodegradable to break down.

15. These are very different questions. There is a very high probability that Professor Frederick would have obtained very similar results had he not shown any pictures of products to survey respondents. The question structure, not the ECM logo, is the likely causal agent with respect to any answers by respondents, and as noted above, the form of the questions, which demand a response in terms of a specific time period, the absence of a basis among respondents

¹⁸ Gordon L. Patzer (1996), *Experiment-Research in Marketing: Types and Applications*, (Westport, CT: Quorum Books), pp. 33–34. See Attachment 15.

¹⁹ For example, the definitions of biodegradation provided by the United States Geological Survey either make no mention of time or note that time is highly variable. U.S. GEOLOGICAL SURVEY, *available at* http://toxics.usgs.gov/definitions/anaerobic_biodegradation.html (last visited June 29, 2015). See Attachment 16.

for providing such an answer, and the falsity of most “acceptable” responses serve to produce meaningless responses to the questions posed in Dr. Frederick’s survey. An “instrument” means a survey, questionnaire, test, scale, rating, or tool designed to measure the variable(s), characteristic(s), or information of interest, often a behavioral or psychological characteristic. As one leading scholar observed: “The instrumentation confound results from changing measurement instruments. Changing the wording of questions in a survey is essentially changing instruments.”²⁰ Researchers cannot derive causal data by comparing two different instruments because researchers cannot know what caused the different responses—it very well could have been the fact that different instruments were used. Without knowing what caused the responses, there can be no causal data.

16. In Section A(2)(b) of their supplemental brief, Complaint Counsel argue that the Synovate survey can provide causal data. Specifically Complaint Counsel argue that Question 8 in the Synovate survey, “How many years do you think it takes for traditional plastic products to biodegrade,” functions as a control for Question 19 in the Synovate survey, “What do you believe is a reasonable amount of time for a ‘biodegradable’ plastic package to decompose in a landfill?” However, no valid causal data can be extracted from the Synovate survey.

17. The Synovate survey failed to define “traditional plastic product.” It is certainly the case that the question did not define “traditional plastic” as non-biodegradable, which is the meaning Complaint Counsel appear to prefer. Question 8 and Question 19 do not contrast “traditional” and “biodegradable” as part of either question. There were also ten questions in between Question 8 and Question 19, which make it highly unlikely any respondent would make

²⁰ Harvey Russell Bernard (2012), *Social Research Methods: Qualitative and Quantitative Approaches*, Second Edition, (Thousand Oaks, CA: Sage), p. 96. See Attachment 17.

a connection between the questions. Finally, Question 8 does not include any information about the environment in which biodegradation occurs while Question 19 is quite specific to landfills. These are yet other examples of an unstated criterion: “If the criteria by which respondents must judge some issue or respond to some question aren’t completely obvious, the criteria must be stated in the question. If an item might be judged by multiple standards and the criteria aren’t explicitly stated, some respondents will use one set of criteria and others will use another.”²¹ This effort to use two very different questions as “test” and “control” is also an illustration of instrumentation error described above.

18. In Section A(2)(c) of their supplemental brief, Complaint Counsel argue that Question 3L in Frederick’s survey can act as a control for Question 4 in the APCO survey. However, that comparison is inappropriate because the questions are not the same. Dr. Frederick’s 3L question asks “how long” while APCO Question 4 asks for “maximum amount of time.” As one leading scholar observed: “The instrumentation confound results from changing measurement instruments. Changing the wording of questions in a survey is essentially changing instruments.”²²

19. The one conclusion that might be drawn across these studies is that representation of a product as biodegradable suggests to consumers that a product will biodegrade faster relative to a product that is not so represented or might not biodegrade at all.

20. In footnote 5 on page 8 of their Supplemental brief, Complaint Counsel state that comparisons can be made across studies based on how the studies frame questions. This is an admission that how questions are asked and “framed” influences response. Ask about weeks,

²¹ See Settle & Alreck, *supra* n. 1 at p. 95. See Attachment 8

²² See Bernard, *supra* n. 20 at p. 96. See Attachment 17.

people respond in weeks; ask about months, people respond in months. Ask without specifying criteria, who knows.

21. In Section B of their supplemental brief, Complaint Counsel argue that the “observational surveys (APCO, Synovate, Dr. Stewart)” corroborate what the experimental studies demonstrate. This misrepresents the survey I conducted which showed unambiguous and near universal understanding of the contextual influences on biodegradation (98%). This is the value of my survey, which is the only one of the four that actually addresses the relevant question in an objective manner.

22. In Section C of their supplemental brief, Complaint Counsel argue that all four studies yield “qualitatively similar result[s].” There is no convergence among the four studies. For the many reasons described in this and earlier reports, the Frederick survey is useless. It was poorly executed, used a suspect sample, asked biased questions, made selective use of data, and cannot be used, consistent with accepted survey research, for any conclusion related to representations by ECM. See above and my earlier report, testimony, and declaration. For reasons already identified in previous reports by both Dr. Frederick and me, the APCO and Synovate studies are of limited value.²³ Misrepresentation of the results of my survey does not demonstrate convergence. Dr. Frederick’s analyses ignore the single most probative extrinsic evidence in this matter: 98% of consumers understand that there is no single, universal temporal period that defines how fast a product will decompose. Consumers as a group exhibit a heterogeneous and diverse understanding of “biodegradation.”

²³ Slavin RE (1986). “Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews,” *Educational Researcher* 15 (9): 5–9 (“a good meta-analysis of badly designed studies will still result in bad statistics”). See Attachment 18.

David W. Stewart

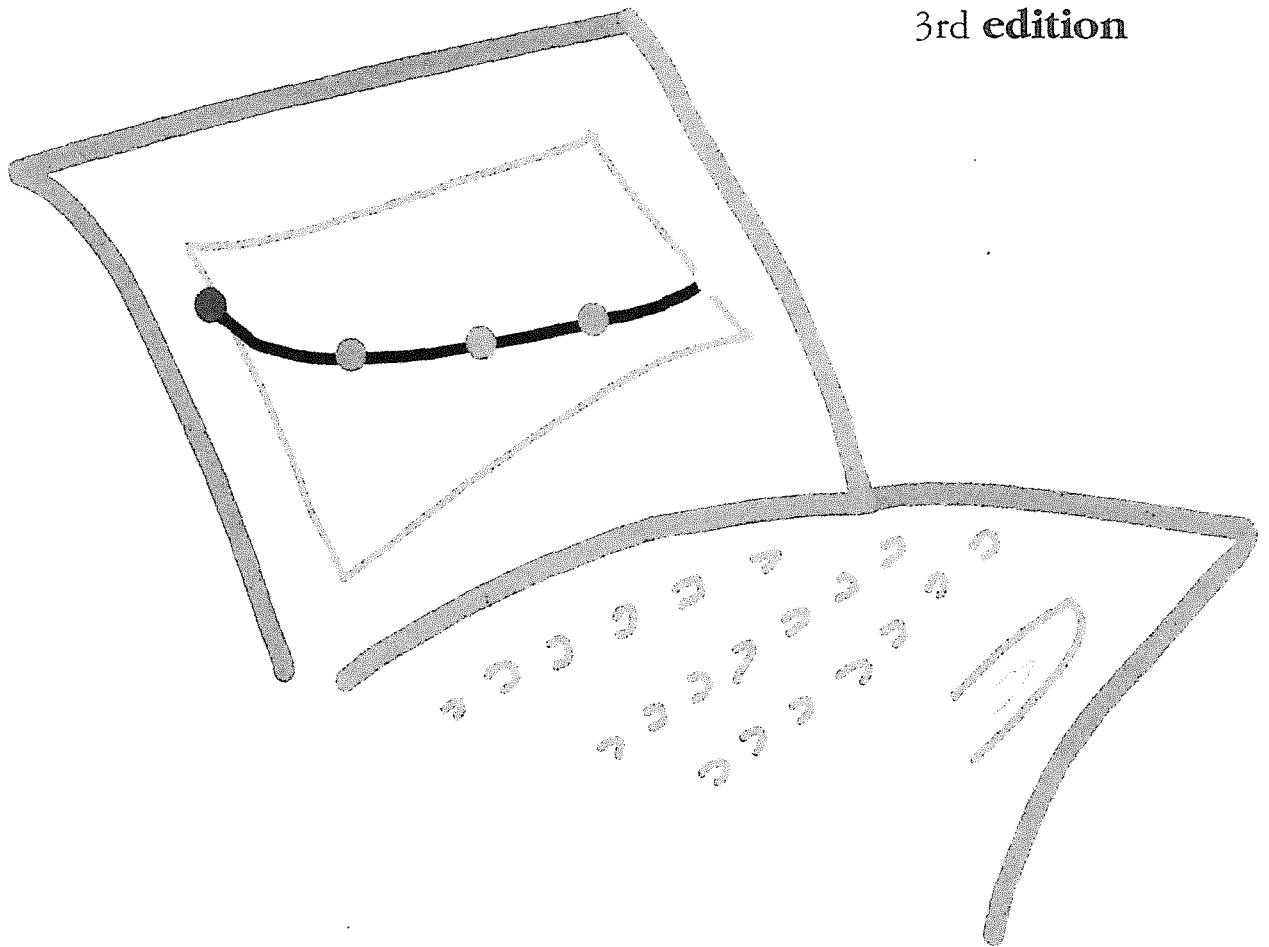
Dr. David W. Stewart

Executed on: July 6, 2015

ATTACHMENT 1

the **Survey** research
h a n d b o o k

3rd edition



pamela l. alreck robert b. **settle**

For Using Unstructured Questions**Guidelist 4-2**

1. Structure the question whenever it's possible to do so, even though it requires time and effort.
2. Resist requests by sponsors for verbatim responses, explaining the difficulty of interpretation.
3. Be sure the dimension or range of alternatives is crystal clear to respondents when an item is unstructured.
4. Be sure the interviewers or respondents can record verbatim responses to unstructured items accurately.
5. Estimate the degree to which using an unstructured question will increase the response task.
6. After all is said and done, go back and see if it may not be possible to use a structured item, anyhow.

answers and comments because the number and volume of responses are much smaller. The advantages of this approach and the recommended procedures are discussed in Appendix A. If the focus groups are conducted *prior* to the survey, the results may be helpful in building the questionnaire and doing the other survey tasks. If the focus group discussions take place *after* the survey has been completed, participants can be questioned about those areas of the survey findings that appear somewhat ambiguous or would benefit from extension and clarification.

Composing Categorical Items

When a structured question is used, the researcher has to choose the categories or response alternatives to be used by respondents. Questions of this kind are called "categorical" items because all responses must fall into a particular category. (Alternative forms of structured items using numeric scales are presented in Chapter 5.) It takes considerable time and effort to compose a categorical question and select the proper categories. On the other hand, if the task is done carefully and thoroughly, it will save a great deal of time and effort later and increase the reliability and validity of the data. Categorical items ask a question, followed by a series of alternative answers. When composing the question itself, the researcher applies all the principles and guidelines discussed earlier. Thus, the same things are required of the question itself, whether it's structured or unstructured.

An All-inclusive List

The categories used with a structured item form a classification system. There are three rules or principles to observe when choosing the categories for such an item: (1) the list must be all-inclusive, (2) the categories must be mutually exclusive, and (3) there should be more variance in the meaning *between* categories than within them.

The first rule is that the list must include every possible response. Every answer a respondent might possibly give must fit into a category, and there must be no conceivable answer that doesn't fit into a category.

Structured Category Questions

Example 4-18

Incorrect Classification

Q. How did you *first* learn about the new clinic?

- A1. From a friend or co-worker.
- A2. From a relative or family member.
- A3. From a newspaper or magazine.
- A4. From the radio or television.
- A5. From a news story.
- A6. By seeing a sign, billboard, or poster.
- A7. By some other announcement or advertisement.

Correct Classification

Q. How did you *first* learn about the new clinic?

- A1. From a relative or family member.
- A2. From an associate or acquaintance.
- A3. From a newspaper or magazine *advertisement*.
- A4. From a radio or television *advertisement*.
- A5. *Read* a news story in some publication about it.
- A6. *Heard* a news story about it on radio or TV.
- A7. Some other way. *Specify how:* _____

In Example 4-18, the incorrect classification scheme shown in the top section doesn't meet the requirement of an all-inclusive set of categories. Suppose some respondents had actually seen the new clinic building. There's no category for recording that means of learning about the clinic. By contrast, the correct version shown in the lower section of the example includes an "other" category, so those responses that don't fit into any of the first six categories can be recorded and identified in the seventh. It's very advisable to include an open "other" category in such lists. Even though the questions are pretested in a pilot survey, there are likely to be a few exceptional or unusual responses that won't fit into any category listed.

When an "other" category is used, the nature of the "other" may or may not be specified, depending on the information requirements. There are times when those seeking information may be interested *only* in a few certain categories, but answers might range widely beyond them. When that's the case, the answers that don't fit into the categories of interest may all be lumped into an "other" category without specification of just what the other things are. For some situations, the nature of the "other" responses may be useful or required by those seeking the information. When they are, they should be specified so they can be identified, categorized, and postcoded when the completed questionnaires or online data files are being edited.

ATTACHMENT 2



MARKETING RESEARCH

Methodological Foundations

GILBERT A. CHURCHILL, JR.
DAWN IACOBUCCI

Ninth Edition

The TV-set purchase question (Table 9.1) illustrates other problems associated with multiple-choice questions. First, the list of reasons cited for purchasing a Sony color TV may not be *exhaustive*. The "other" response category attempts to solve this problem. However, if many respondents check the "other" category, the study will be useless. Thus, the burden is on the researcher to make the list of alternatives comprehensive.

Unless the respondent is instructed to check all alternatives that apply, or is to rank the alternatives in order of importance, the multiple-choice question also demands that the alternatives be *mutually exclusive*. The income categories shown below violate this principle:

- \$10,000–\$20,000
- \$20,000–\$30,000

A respondent with an income of \$20,000 would not know which alternative to check. When questions are about a product's features, (a TV's picture, warranty, price, etc.), you often see instructions such as, "Check the most important reason," "Check all those reasons that apply," or "Rank all the reasons that apply from most important to least important."

The list of alternative responses should be *exhaustive*, but a long list will be *exhausting* to respondents! So, when designing multiple-choice questions, the researcher should remain cognizant of human beings' limited data-processing capabilities.

The fourth weakness of the TV-purchase question is that it may be susceptible to order bias. The recommended procedure for combating this order bias is to prepare several forms of the questionnaire, with several different orders. If each alternative appears early, late, and in the middle across the different forms, the researcher can feel reasonably comfortable that position bias has been neutralized.⁷

The long-distance telephone call example in Table 9.1 illustrates a problem with questions designed to get at the frequency of behaviors. The range of the categories used in the question seems to cue respondents about how they should reply. A scale with the following categories would likely produce a different picture than the one shown in Table 9.1:

- Fewer than 10
- 10–20
- More than 20

A financial institution has developed a new type of savings bond. The marketing director of this institution has requested that a local research supply company design a questionnaire that will help quantify target consumers' interest in this new bond. However, the marketing director is concerned about the possibility that competitors will hear about the new product concept because of the survey. He requests that the questionnaire be written in such a way that the true purpose of the study is masked.

To mask the actual purpose of the study, the questionnaire primarily asks respondents for details of their holiday plans and budgets. Because respondents are asked questions about their finances only after being asked multiple vacation-related questions, it is hoped that respondents will assume the information is for a travel company. Moreover, the marketing director of the financial institution asks that interviewers tell respondents that the information is being gathered for a travel-related company.

- Discuss the implications of deceiving respondents on a questionnaire in this way.
- If the interviewers had not been told to explicitly tell respondents that the information was for a travel-related corporation, would the deception be acceptable?
- Are there ways of acquiring this type of information without resorting to deception while still protecting the institution's new product idea?
- Discuss the validity issues associated with respondents knowing the purpose of the survey as they are completing it.

⁷ David Moore, "Measuring New Types of Question-Order Effects," *Public Opinion Quarterly* 66 (2002), pp. 80–91.

ATTACHMENT 3

Encyclopedia of Survey Research Methods

$$DEFF_R = \frac{V_{CSD}(y)}{V_{SRS}(y)}$$

$$fpe = \frac{N-n}{N} = 1 - f$$

$$\sum_{i=1}^C \sum_{j=1}^M \sum_{k \neq l}^M (y_{ij} - \bar{y}_{i0})(y_{ik} - \bar{y}_{i0})$$

EDITOR

Paul J. Lavrakas

1

VOLUME

to repair misunderstandings and to clarify question objectives, or some combination of both retrieval cuing and conversational interviewing. Ongoing verbal behavior coding studies that document the occurrence of different types of retrieval cues and conversational mechanisms may uncover which types of verbal behaviors produce better data quality. Such work is likely to lead to improvements in interviewer training.

That event history calendars show mostly encouraging gains in data quality in comparison to standardized interviewing indicates that it is not a panacea that will "cure" all ills associated with forgetting, and that there are also likely beneficial aspects to standardization that are not utilized in event history calendar interviews. The very few studies that have been conducted have shown that event history calendar interviewing leads to modest increases in interviewer variance in most, but not all, instances. The event history calendar also usually leads to modest increases in interviewing time, at present on the order of 0%–10% longer than standardized interviews. Interviewers show overwhelming preference for event history calendar interviewing in ease of administration. As an attempt to acquire the "best of both worlds," hybrid event history calendar and standardized interviewing instruments have also been designed.

Administration Methods

Event history calendars have been administered in a variety of methods, including as paper-and-pencil and computer-assisted interviewing instruments, and in face-to-face, telephone, and self-administered modes. The method has mostly been implemented in the interviewing of individuals, but the interviewing of collaborative groups has also been done. The computerization of event history calendars affords the automation of completeness and consistency checks. Web-based applications are also being explored.

Robert F. Belli and Mario Callegaro

See also Aided Recall; Conversational Interviewing; Diary; Interviewer Variance; Reference Period; Standardized Survey Interviewing

Further Readings

Axinn, W. G., & Pearce, L. D. (2006). *Mixed method data collection strategies*. Cambridge, UK: Cambridge University Press.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, 6, 383–406.

Belli, R. F., Shay, W. L., & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public Opinion Quarterly*, 65, 45–74.

Belli, R. F., Stafford, F. P., & Alwin, D. F. (in press). *Calendar and time diary methods in life course research*. Thousand Oaks, CA: Sage.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology*, 18, 37–68.

Yoshihama, M., Gillespie, B., Hammock, A. C., Belli, R. F., & Tolman, R. M. (2005). Does the life history calendar method facilitate the recall of intimate partner violence? Comparison of two methods of data collection. *Social Work Research*, 29, 151–163.

EXHAUSTIVE

Exhaustive is defined as a property or attribute of survey questions in which all possible responses are captured by the response options made available, either explicitly or implicitly, to a respondent. Good survey questions elicit responses that are both valid and reliable measures of the construct under study. Not only do the questions need to be clear, but the response options must also provide the respondent with clear and complete choices about where to place his or her answer. Closed-ended or forced choice questions are often used to ensure that respondents understand what a question is asking of them. In order for these question types to be useful, the response categories must be mutually exclusive and exhaustive. That is, respondents must be given all possible options, and the options cannot overlap. Consider the following question, which is frequently used in a number of different contexts.

Please describe your marital status. Are you . . .

Married

Divorced

Widowed

Separated

Never married

This question does not provide a response option for couples who are in committed relationships but are not married, whether by choice or because of legal barriers. For example, a woman who has been with a female partner for 5 years would be forced to choose either married or never married, neither of which accurately describes her life situation. Without a response option that reflects their life circumstances, those respondents may be less likely to complete the questionnaire, thus becoming nonrespondents. This question is easily improved by the addition of another response category:

A member of an unmarried couple

In situations in which the researcher cannot possibly identify all response options a priori, or cannot assume a single frame of reference for the subject matter, an "Other [specify]" option can be added. For example, questions about religion and race always should include an "Other [specify]" option. In the case of religion, there are too many response options to list. For race, traditional measures often do not adequately capture the variety of ways in which respondents conceptualize race. Thus, an "Other [specify]" option allows respondents to describe their race in a way that is most accurate to them.

Linda Owens

See also Closed-Ended Question; Forced Choice; Mutually Exclusive; Open-Ended Question

Further Readings

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

EXIT POLLS

Exit polls are in-person surveys in which data are gathered immediately after people have engaged in the behavior about which they are being surveyed, such as voting in an election. The survey methods that are used in exit polls apply to the measurement of a wide variety of behaviors, but in the minds of most people exit polls are most closely associated with what is done on Election Day to help project the

winning candidates before the final vote tally is announced. Although political exit polling is done in many countries, it is the exit polling conducted for elections in the United States that is covered here.

How Exit Polling Is Conducted and Used in U.S. Elections

The exit polls that are conducted nationwide and in most individual states for the general election in the United States are among the largest single-day surveys that are conducted anywhere, with data from more than 100,000 respondents being gathered, processed, and analyzed within one 24-hour period.


To estimate the outcome of an election in a particular geopolitical area of the United States, which most typically is done at the state level, a stratified random sample of voting precincts within the area is selected, and at least one interviewer is sent to each of the sampled precincts. In the 2004 U.S. general election, there were 1,469 sampled precincts nationwide, and in 2006 there were 993. Those exit polls were conducted by Edition Media Research and Mitofsky International, the organizations that were hired to gather the exit poll data for their news media funders (ABC, the Associated Press [AP], CBS, CNN, Fox, and NBC). On a systematic basis, and in order to obtain a completed questionnaire, the exit poll interviewer stops (i.e., intercepts) people who just finished voting as they exit from their voting places. For example, the interviewers may do this with every 10th person who comes out of the voting place. In each sampled precinct, an average of approximately 100 voters is interviewed over the course of Election Day. Not all exiting voters who are stopped agree to complete the exit poll questionnaire, but in those cases the interviewer records basic demographic information about these refusing voters. This information is used later as part of analyses that investigate the nature of exit poll nonresponse. Interviewers at each sampled precinct telephone in the data they are gathering at three scheduled times on Election Day: mid-morning, early afternoon, and within the hour before voting ends in the precinct.

In order to gather the exit poll data, the interviewer typically hands the selected voter a questionnaire on a clipboard and asks her or him to complete it and then deposit it in a survey "ballot box." The questionnaire gathers three types of data: (1) it measures who

ATTACHMENT 4

Howard Schuman
Stanley Presser

Questions & Answers in Attitude Surveys



Experiments On Question Form,
Wording, And Context

this book is the extent to which people, once they have agreed to be interviewed, accept the framework of questions and try earnestly to work within that framework. If we do not provide a particular substantive alternative to a closed question, people rarely give it. If we omit a don't know category or a middle alternative, people ordinarily do not volunteer one—let alone insist on it. Question constraints are not absolute, and in extreme situations, as when (in Chapter 5) we ask people questions about objects they have never heard of, the majority will rebel. But for most questions people accept the "rules of the game," as they are conveyed by the form of the question.

QUESTION CONSTRAINT AND RESPONSE PERSUASION

The concept of *question constraint* provides a useful starting point in attempting to understand the effects of question form on response marginals. The largest of these effects can be viewed as a result of the omission of categories that many respondents would like to use if they are available. This is most obvious for two of the forms we studied: don't know (DK) filters and middle alternatives (MA). For the former, the choice of don't know is higher by an average (median across items) of about 22% when the don't know alternative is read to respondents; in the latter case, the choice of a middle alternative increases by an average of about 15% when it is offered. It seems sensible to regard these increments as representing respondent preferences and to assume that it is the *decreases* on the versions that omit such alternatives that represent artificial constraint by question form. This is not to argue that the constraint is unreasonable from the standpoint of the investigator's goals, but only that it is indeed a constraint. Moreover, these estimates are minimal ones, for our experiments always included instructions to interviewers to accept DK and MA responses when they were given spontaneously: Without such instructions the form differences would certainly be greater, although by how much we do not know.

Question constraint plays an equally strong role in open-closed question comparisons, but its effects there are double-edged. The most obvious impact is that responses on a closed question are largely limited to the substantive alternatives listed, whereas an open version of the same question produces a much wider array of responses. Even where we took great pains to maximize comparability between the main open codes and closed alternatives, a noticeable proportion of open responses

fell outside the codes common to the two forms. Yet, paradoxically, the completely open form of a question can also be constraining, even though in theory almost any response to it is acceptable. This is because the question itself inevitably implies a frame of reference, and this frame can be defined more broadly for respondents by listing alternatives that they did not realize were legitimate (see Chapter 3, p. 85).

In a sense even the absence of attitude strength measures in a questionnaire is constraining, since this prevents respondents from indicating qualifications to a response and thus forces answers to sound more black and white than they often are.

Only in the case of the balance problem are we unsure as to whether question constraint as such plays a role. If what we have called formal balance (p. 180) increased "no" responses appreciably, one might argue that the absence of an explicit negative alternative in unbalanced questions is constraining. However, if one regards the addition of "or oppose" in a question as similar to the addition of "don't know" or a middle alternative, then our evidence is that the former has virtually no effect, whereas the latter, as already indicated, have substantial impact. Thus unbalanced questions apparently imply the omitted negative category so clearly that its absence from the wording of the question imposes no constraint. The further step of adding a counterargument, which does have an effect, seems better seen as introducing a second force, that of persuasion, into the question-answer process. The counterargument carries an informational, logical, or emotional message that influences the respondent toward a different answer. Although this might also be considered a kind of constraining pressure, it is rather different from the previous cases, which were based on defining alternatives as within or outside the question frame of reference.

Agree-disagree statements, however, raise a further complication, for one way of conceptualizing the phenomenon labeled acquiescence is that some respondents feel constrained to agree with an assertion offered by the interviewer. In that case the substitution of a forced-choice form can be seen as freeing respondents from constraints imposed by the joint operation of the agree-disagree form and an acquiescent tendency by the individual. Moreover, since in two experiments we found that interrogative forms are not less acquiescence-prone than statements, the same argument can be made for all one-sided questions. In sum, even where a purely formal negative alternative (disagree) is offered to respondents, there *may* still be constraint imposed by the affirmative thrust of the question.

The resolution of this issue is difficult but important. It is clear that

question constraint plays a critical role in defining what is a legitimate response in a simple "rules of the game" or definition of the situation sense. We believe that persuasion of a more active sort also affects responses, but the evidence is not quite so self-evident in interpretation.¹ And the mixture of the two processes—of question constraint and response persuasion—is even less well understood. A direct attempt to distinguish the two processes would be valuable.

Primary and Secondary Effects on Marginals

There is one further issue having to do with marginal distributions that is peculiar to don't know and middle-alternative experiments: The effect on the main substantive alternatives of the form variation. For don't know and middle alternative experiments, the response that is varied is usually not the primary concern of the research, and it is possible for its proportion to change substantially without altering the *relative* distribution of other responses. In fact, this is exactly what occurs: For most don't know and middle alternative comparisons, the relative proportions choosing the substantive alternatives (e.g., pro and con on an issue) are much the same on both question forms, although there are a few exceptions that prevent complete generalization.

The situation is different for the other types of question form variations that we studied. In these cases the variation directly involves a substantive alternative, hence an effect produces by its very nature a change in substantive marginals. If a counterargument is added to a question, or an agree-disagree item transformed into a forced-choice item, any effect that occurs has an immediate impact on the primary response marginals.

These considerations lead to some qualification of the common belief among experienced survey researchers that almost any change in question wording will affect question marginals. For although inclusion of a don't know or middle alternative certainly changes the proportion of persons taking that choice—often shifting more than a quarter of the sample—there is usually no other detectable change in question margi-

¹Under unintended response persuasion one might include presumed effects due to variation in the social desirability of closed alternatives. But here also the evidence is not strong, for our attempts in Chapter 3 to isolate such an artifact were unsuccessful. More positive evidence on persuasion due to wording effects appears in Chapter 11, where the nonsubstantive addition of the word "freedom" to a question has an impact that seems to involve a kind of persuasion.

ATTACHMENT 5

VOLUME ONE

**THE HANDBOOK OF
SOCIAL PSYCHOLOGY**

FOURTH EDITION

**DANIEL T. GILBERT
SUSAN T. FISKE
GARDNER LINDZEY**

cluded that American tolerance for dissent had increased significantly over the intervening decades. This finding reflects a common assumption and typical finding in survey research: correlational results are less sensitive to response effects than are single item distributions or "marginals" (Stouffer & DeVinney, 1949). This assumption is not without exceptions, however, as later examples will indicate.

Question Constraint The observation that respondents ordinarily accept the framework provided by a question and try to answer within it is referred to as *question constraint*. Although the concept of question constraint has typically been applied to response alternatives, it is worth noting that every question imposes a perspective that is usually taken for granted by respondents (see Clark & Schober, 1992). For example, in 1980, British respondents were asked, "The government believes that in the interest of fighting inflation, local authority [municipal] workers should get no more than a 6 percent pay raise this time. Do you agree with this view or do you think they should get more?" As Turner and Martin (1984, p. 80) noted, this question introduces the presupposition that a ceiling on increases in pay will fight inflation, yet this presupposition is not the subject matter of the question. However, respondents are likely to work within the constraints of the question presented to them, selecting one of the response alternatives rather than questioning the entailed presupposition. Researchers therefore need to keep in mind that it is they, not the respondents, who are determining the framework within which survey answers are given.

Respondents' tendency to work within the constraints imposed by the question is particularly apparent with regard to "don't know" (DK) or "no opinion" responses. Standard survey questions usually omit "no opinion" as an explicit option, but instruct interviewers to accept this response when volunteered. Experimental studies (e.g., Bishop, Oldendick, & Tuchfarber, 1983; Schuman & Presser, 1981) have consistently found that respondents are more likely to report not having an opinion when a DK option is explicitly offered (see Schwarz & Hippler, 1991, for a review). Similarly, many respondents prefer a *middle alternative* between two extreme positions when offered, but endorse one of the extremes when the middle alternative is omitted (e.g., Bishop, 1987; Kalton, Collins, & Brook, 1980; Schuman & Presser, 1981). Thus, most respondents assume that the rules of the game call for working within the categories offered, even though a desire to answer otherwise is evident when more choice is provided. As Clark and Schober (1992) noted, however, acceptance of the constraints imposed by the questioner is not unique to survey interviews but characterizes question answering in more natural contexts as well, although its effects may be less apparent in daily conversations.

The extent to which question constraint can affect an-

swers has been demonstrated in two experiments on open and closed questions (Schuman & Scott, 1987). The first experiment showed that when a widely used open-ended question about "the most important problem facing this country today" was converted into a closed question listing four specific problems, the listed responses rose dramatically ("quality of public schools" increased from 1 percent to 32 percent), while almost none of the common responses to the open question (e.g., "unemployment") were offered much despite the encouragement for "other" answers. Lest this suggest that the solution to avoiding question constraint is to ask open questions, a second experiment showed that an open inquiry about important events of the past half century elicited only a few mentions of "the invention of the computer," but when the invention of the computer was included as an alternative in a closed question, it was the most frequently chosen answer, exceeding even World War II which had been the leading mention to the open question. In this case, other evidence suggested that the computer response was indeed highly important to respondents, but that the open question had unintentionally limited the scope of thinking to events of a political nature. Thus, there is no purely formal way to avoid question constraint entirely in survey questioning, and investigators need to constantly be aware of the limits they are themselves imposing on their respondents.

Summary Our selective discussion of issues of question wording indicates that apparently minor changes in the specific wording of the question stem, or in the response alternatives presented, may have a pronounced impact on the obtained responses. Whereas some observers concluded from such findings that respondents provide relatively meaningless answers, or else their opinions wouldn't change as a function of minor wording changes (e.g., Crossen, 1994), we prefer a more optimistic summary. In our reading, these findings indicate that respondents pay close attention to the question asked, treating the specifics of the wording and the response alternatives offered as relevant contributions to the ongoing conversation. As noted earlier, this is, indeed, what they are entitled to on the basis of the tacit norms that govern the conduct of conversation. Moreover, respondents draw on these specifics in their efforts to infer the meaning intended by the questioner, much as the tacit rules of conversational conduct would want them to (see Schwarz, 1994, 1996, for a more detailed discussion). Hence, the apparent artifacts of question wording are likely to reflect regularities of normal conversational conduct, except that we as researchers often fail to take these regularities into account in writing questions and interpreting answers.

Question Order: The Emergence of Context Effects
Survey researchers have long been aware that the order in

ATTACHMENT 6



Problems in the Use of Survey Questions to Measure Public Opinion

Howard Schuman; Jacqueline Scott

Science, New Series, Vol. 236, No. 4804 (May 22, 1987), 957-959.

Stable URL:

<http://links.jstor.org/sici?sici=0036-8075%2819870522%293%3A236%3A4804%3C957%3APITUOS%3E2.0.CO%3B2-Z>

Science is currently published by American Association for the Advancement of Science.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/aaas.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

REFERENCES AND NOTES

1. D. Picard and W. Schaffner, *Nature (London)* 307, 80 (1984); E. B. Reilly, unpublished results.
2. K. Shimotohno and H. M. Temin, *Cell* 26, 67 (1981); M. Emerman and H. M. Temin, *J. Virol.* 50, 42 (1984); A. L. Joyner and A. Bernstein, *Mol. Cell. Biol.* 3, 2180 (1983).
3. C. L. Cepko *et al.*, *Cell* 37, 1053 (1984).
4. R. Mann, R. C. Mulligan, D. Baltimore, *ibid.* 33, 153 (1983).
5. J. G. Izant and H. Weintraub, *ibid.* 36, 1007 (1984); S. K. Kim and B. J. Wold, *ibid.* 42, 129 (1985).
6. R. Cone *et al.*, *Mol. Cell. Biol.* 7, 887 (1987).
7. O. Bernard, N. Hozumi, S. Tonegawa, *Cell* 15, 1133 (1978).
8. G. J. Todaro and H. Green, *J. Cell Biol.* 17, 299 (1963).
9. E. J. Siden, D. Baltimore, D. Clark, N. E. Rosenberg, *Cell* 16, 389 (1979).
10. C. J. Paige, P. W. Kincaide, P. J. Ralph, *J. Immunol.* 121, 641 (1978).
11. A. W. Harris, A. D. Bankhurst, S. Mason, N. L. Warner, *ibid.* 110, 431 (1973).
12. K. Horibata and A. W. Harris, *Exp. Cell Res.* 60, 61 (1970).
13. J. M. Chirgwin *et al.*, *Biochemistry* 18, 5294 (1979).
14. E. B. Reilly, A. R. Frackelton, Jr., H. N. Eisen, *Eur. J. Immunol.* 12, 552 (1982).
15. M. Reith, T. Imanishi-Kari, K. Rajewsky, *ibid.* 9, 1004 (1979).
16. U. K. Laemmli, *Nature (London)* 227, 680 (1970).
17. W. M. Bonner and R. A. Laskey, *Eur. J. Biochem.* 46, 83 (1974).
18. D. Hausteil, J. J. Marchalonis, A. Harris, *Biochemistry* 12, 1130 (1973).
19. The 7.4-kb Eco RI fragment encoding the rearranged $\lambda 1$ gene from HOPC2020 was a gift from B. Blomberg; monoclonal antibody to $\text{C}\lambda 1$ (LS 136) was a gift from T. Imanishi-Kari. Supported by grants from NIH (CA26712 and CA38497) and the MacArthur Foundation to R.C.M.

13 August 1986; accepted 26 February 1987

Problems in the Use of Survey Questions to Measure Public Opinion

HOWARD SCHUMAN AND JACQUELINE SCOTT

Sample interview surveys are frequently proposed and sometimes used as a way of studying public choices among alternatives. Questions in such surveys may be either "open" or "closed." Two experiments are reported that demonstrate the difficulty of inferring not only absolute levels but even relative orderings of public choices from either type of question, although such questions can be used more successfully to study temporal change or variations across social categories.

A SEEMINGLY SIMPLE WAY OF ASSESSING public opinion is to ask a random sample of the public to choose from among an explicit or implicit set of alternatives. The form of the question, however, can greatly affect such choices. One crucial distinction is whether respondents are expected to answer in their own words from alternatives they construct (open questions) or to select instead from a list of offered alternatives (closed questions). Very little research has been carried out on what effect this difference in question form makes in studying public opinion (1).

We present experimental evidence on the limitations of both open and closed questions in attempts to measure public choices. Closed questions are shown to sharply restrict frames of reference by focusing attention on the alternatives offered, no matter how impoverished those alternatives may be and no matter how much effort is made to offer respondents freedom to depart from them. Open questions are shown to exercise their own form of constraint, though in subtle ways that can easily be missed by investigators. The goal of the experiments is not to argue against either form of question, but to emphasize that question content is always based, whether recognized or not, on important assumptions about what should be included in respondent frames of reference. The unexamined question is not worth asking.

Limitations of closed questions. For this experiment we employed a frequently used open question, that about "the most impor-

tant problem facing this country today" (Table 1) (2). This open question was asked to a random half of a national sample in the October 1986 Monthly Random Digit Dial Telephone Survey conducted by the Survey Research Center. The other half of the sample was asked a specially constructed closed version of the question (Table 1). The closed version listed four problems, each of which had been mentioned by less than 1% of the population in recent use of the open question by the Gallup organization. Respondents were not, however, forced to choose one of these rare alternatives, but were told as part of the question that "if you prefer, you may name a different problem as most important."

As expected, Table 1 shows that less than 3% of the national sample spontaneously mentioned any of the four "rare" problems to the open question. The categories most frequently coded were unemployment (17%), general economic problems (17%), threat of nuclear war (12%), and foreign affairs (10%), with the rest of the responses scattered among a dozen categories, including 5% "don't know."

On the closed form, however, 60% of the sample chose one of the four "rare" alternatives as "most important," only 40% taking the option of naming some other problem. Moreover, unemployment, the most frequently mentioned single problem on the open form, was given by only 6.2% of the respondents on the closed form.

On the basis of the closed question, one would conclude that the quality of public

schools is what troubles Americans most, followed by the issue of pollution and then by abortion, whereas on the open question it is economic and international problems that loom largest, while the issues of education, pollution, and abortion are practically invisible.

Most readers will assume, as do we, that the issues mentioned on the open question give the better overall picture of American concerns and that the findings on the closed question are distorted by the constraint or inertia produced by listing the four problems as part of the question, despite the explicit provision offered to respondents to depart from them (3).

The limitations of open questions. The preceding results suggest that open questions provide a clearer picture of the concerns of a survey sample than do closed questions. Yet this ignores the possibility that open questions can also provide a constraining frame of reference. The following experiment was carried out to test this assumption as clearly as possible.

The experiment was suggested in the course of another survey. Respondents had been asked to name one or two of the most important "national or world event (events) or change (changes)" during the past 50 years that came to mind. To this open question, the most commonly given responses had to do with World War II and the Vietnam War, but, as intended, many answers referred to broader social changes, such as the civil rights movement or to scientific and technological developments, such as space exploration. Hardly mentioned at all, however, was the development of the computer, which might not have seemed surprising except that references to computers occurred frequently in responses to later questions.

This discrepancy suggested that computers had made a considerable impact on the public, but that the "national or world event or change" open question tended unwittingly

Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106-1248.

Table 1. Offering rare responses. The open question was, "What do you think is the most important problem facing this country today?" The closed question was, "Which of the following do you think is the most important problem facing this country today—the energy shortage, the quality of public schools, legalized abortion, or pollution—or if you prefer, you may name a different problem as most important."

Category	Response (%)	
	Open question	Closed question
The energy shortage	0.0	5.6
The quality of public schools	1.2	32.0
Legalized abortion	0.0	8.4
Pollution	1.2	14.0
All other responses	93.0	39.3
Don't know	4.7	0.6
Total	100 (<i>n</i> = 171)	100 (<i>n</i> = 178)

ly to preclude such responses. The question probably focuses thoughts on the broad political domain, and even where this is not the case, as with space exploration responses, dramatic incidents like the moon landing may have been necessary to yield a large response category.

To test the hypothesis that computer-related responses would be much more common if offered as an explicit choice, we included "the invention of the computer" as an alternative to a closed question, as shown in Table 2, along with the four categories that had been most commonly given to the open question in the earlier study. This closed question was administered to a random half of a national monthly telephone sample in July and August 1986, the other half receiving an exact replication of the original open question (4).

It must be emphasized that the closed question in this case was quite differently constructed than the closed "most important problem" question in Table 1, since the latter had offered only rarely given open categories. In the present instance, exactly

the opposite was done: with the exception of the computer category, the alternatives offered as part of the closed question were based on the most frequently given open responses. Previous research indicates that such categories will increase in size when read as part of a closed question, but that their ranking will not change appreciably relative to one another (5). In the present experiment, however, we hypothesized that "the invention of the computer" would increase in frequency much more than other alternatives when offered as part of a closed question, because many respondents would realize that it is a legitimate response to the question. The data in Table 2 strongly support this hypothesis: on the open question, the invention of the computer is the least frequently mentioned (1.4%) of the five categories that are our concern; on the closed question it is given by 30% of the public, becoming the modal response for the entire sample. Moreover, the statistically significant likelihood-ratio chi square for the question form by five category table ($\chi^2 = 54.2$, *df* = 4, *P* < 0.001) is based almost

entirely on the contrast of the computer response with all others ($\chi^2 = 50.3$, *df* = 1); when that row is omitted the remaining two by four table no longer approaches significance ($\chi^2 = 3.9$, *df* = 3).

There is further evidence that the process revealed in Table 2 is different from that in Table 1. In the previous experiment, despite the impressive constraint produced by the listing of rare categories in the closed question, nearly 40% of the sample did choose to go outside the listed alternatives. In the present experiment virtually everyone (94%) was satisfied to select one of the listed choices, which is consistent with our having deliberately included the four that are most frequently given spontaneously plus one that we hypothesized to be a potentially preferred choice once it is made legitimate and equally salient. It is also noteworthy that the space exploration category did not show a jump similar to the computer alternative on the closed form, indicating that more than a shift of emphasis from political to nonpolitical answers was involved. More likely it was a shift from changes that reach public consciousness through dramatic incidents (for example, the televised moon landing) to changes that are more gradual and cumulative in impact (the computer) (6).

These two experiments demonstrate how misleading univariate distributions can be in representing public choices. On the one hand, respondents tend to choose among the alternatives offered to them, even where they are explicitly instructed that this is not necessary. If an investigator wishes to know how the public ranks all alternatives that come to mind, the initial ranking must be provided in a free answer situation. This in itself is not an insurmountable problem, since it is possible to proceed in a two-step sequence: first, obtain spontaneous expressions by the public, then use these to construct a set of closed choices (7).

However, this strategy assumes that answers provided to an initial open question do represent what respondents have "in mind." This may be the case in terms of the respondent's interpretation of the wording of the open question, but not in terms of the investigator's goals. Our second experiment showed that the wording of the open question may constrain respondents by not legitimizing types of responses that the investigator had intended to include. There is no simple way around such a constraint, since investigators themselves are likely to be unaware that respondents are unaware of the possibility of giving such responses. In studies that are attempting to determine frames of reference, there is no substitute for repeated efforts to learn in a variety of loosely

Table 2. Omitting possible responses. The open question was, "There have been a lot of national and world events and changes over the past 50 years—say from 1930 right up until today. Would you mention one or two such events or changes that seem to you to have been especially important. There aren't any right or wrong answers to the question—just whatever national or world event or change over the past 50 years that comes to mind as important to you." The closed question was, "There have been a lot of national and world events and changes over the past 50 years—say from about 1930 right up until today. Would you choose from the list I read the event or change that seems to you to have been the most important, or if you wish you can name an event or change different from the ones I mention. There aren't any right or wrong answers to the question—just whatever national or world event or change over the past 50 years that seems most important to you. Here is the list: World War II, the exploration of space, the assassination of John F. Kennedy, the invention of the computer, or the Vietnam War?"

Category	Response (%)	
	Open question	Closed question
World War II	14.1	22.9
Exploration of space	6.9	15.8
Assassination of John F. Kennedy	4.6	11.6
Invention of the computer	1.4	29.9
The Vietnam War	10.1	14.1
All other responses	52.2	5.4
Don't know	10.6	0.3
Total	100 (<i>n</i> = 347)	100 (<i>n</i> = 354)

6. Although we did not compare the computer response on the two question forms—there are only five such cases on the open form—the closed form does yield a highly significant correlate for the computer response versus all other closed choices. “The invention of the computer” was chosen especially by the youngest (18 to 29) of four age categories. Closer study of the two forms suggests that young people tended to give space-related responses to the open question, but shifted to the computer response on the closed form.

7. There is evidence that such a sequence can produce close correspondence between open and closed question distributions. See Schuman and Presser in (1) and Schuman, Ludwig, and Krosnick (5).

8. Exactly this point was made explicitly in one of the first major uses of survey data [S. A. Stouffer *et al.*, *The American Soldier: Adjustment During Army Life* (Princeton Univ. Press, Princeton, NJ, 1949)]. Of course, all such comparisons assume that the form of the question has been held constant.

9. This report draws on research supported by NSF grants SES-8411371 and SES-8410078. Helpful suggestions on the manuscript were made by P. E. Converse, R. M. Groves, and J. Ludwig.

29 December 1986; accepted 25 March 1987

structured ways what respondents have “on their minds.”

There is one practical solution to the problems pointed to in this report. The solution requires giving up the hope that a question, or even a set of questions, can be used to assess preferences in an absolute sense or even the absolute ranking of preferences and relies instead on describing changes in responses over time and differences across social categories (3). The same applies to all survey questions, including those that seem on their face to provide a picture of public opinion (8).

REFERENCES AND NOTES

1. See S. Sudman and N. M. Bradburn [*Asking Questions* (Jossey-Bass, San Francisco, 1982)] for a discussion of the open-closed distinction. H. Schuman and S. Presser [*Questions and Answers in Attitude Surveys* (Academic Press, New York, 1981)] and a few earlier but marginally relevant reports cited therein provide partial exceptions to the statement about lack of research for questions about public opinion.
2. Although relatively simple questions were used in this investigation in order to provide precision in results, there is little reason to think that the basic conclusions will differ when questions are more complex.
3. We cannot compare correlates of the four focal issues on the two question forms, since there are too few cases in those categories on the open form to allow comparison. Considering the closed form alone, there is no significant relation between education and choosing one of the listed alternatives rather than giving an unlisted problem. However, among the four closed alternatives taken as a set, there are statistically significant associations with education, for example, choice of quality of public schools increases with respondent educational level.
4. The closed question was itself divided into five randomly administered forms, each with a different ordering of the five alternatives. No significant order effect was discovered in this sub-experiment, though such effects occur for some questions; see H. Schuman and S. Presser in (1).
5. H. Schuman, J. Ludwig, J. Krosnick, *Public Opin.*

Lipoprotein Uptake by Neuronal Growth Cones in Vitro

MICHAEL J. IGNATIUS, ERIC M. SHOOTER, ROBERT E. PITAS, ROBERT W. MAHLEY

Macrophages that rapidly enter injured peripheral nerve synthesize and secrete large quantities of apolipoprotein E. This protein may be involved in the redistribution of lipid, including cholesterol released during degeneration, to the regenerating axons. To test this postulate, apolipoprotein E-associated lipid particles released from segments of injured rat sciatic nerve and apolipoprotein E-containing lipoproteins from plasma were used to determine whether sprouting neurites, specifically their growth cones, possessed lipoprotein receptors. Pheochromocytoma (PC12) cells, which can be stimulated to produce neurites in vitro, were used as a model system. Apolipoprotein E-containing lipid particles and lipoproteins, which had been labeled with fluorescent dye, were internalized by the neurites and their growth cones; the unmetabolized dye appeared to be localized to the lysosomes. The rapid rate of accumulation in the growth cones precludes the possibility of orthograde transport of the fluorescent particles from the PC12 cell bodies. Thus, receptor-mediated lipoprotein uptake is performed by the apolipoprotein B,E(LDL) (low density lipoprotein) receptors, and in the regenerating peripheral nerve apolipoprotein E may deliver lipids to the neurites and their growth cones for membrane biosynthesis.

INJURED MAMMALIAN PERIPHERAL nerves can regenerate for long distances through a distal sheath populated by Schwann cells, macrophages, and other non-neuronal “sheath” cells (1). When transplanted into injured central nervous system (CNS) pathways, these peripheral nerve sheaths can support growth of normally nonregenerating CNS axons (2). Attention has therefore been directed at identifying factors present in the injured peripheral nerve that might initiate or facilitate the growth of the damaged fibers (3). One candidate is a soluble protein of M_r 37,000. The rate of synthesis of this protein increases dramatically after injury to an adult rat

sciatic nerve; this protein can account for nearly 5% of the total protein secreted by the nerve 3 weeks after injury (4). This protein, identified as apolipoprotein E (apo-E) (5), is produced by the macrophages that enter the damaged nerve within 3 days of injury (6).

Apolipoprotein E is associated with various plasma lipoproteins, including high density lipoproteins (HDL), and participates in the transport of cholesterol into various cells. It serves as a ligand for the apo-B,E(LDL) (low density lipoprotein) receptor, which mediates the uptake of the apo-E-containing lipoproteins and provides cells with lipids for various metabolic pathways,

including membrane biosynthesis (7). Thus, apo-E may participate in the redistribution of lipids to various cells in neural tissue through similar mechanisms.

After nerve injury, the cholesterol released from myelin membranes is reused in the reassembly of both myelin and axonal membranes in the regenerating nerve (8). It has been suggested (5, 9) that apo-E secreted by macrophages in the injured nerve provides the mechanism for lipid reutilization by facilitating the production of apo-E-containing lipoproteins that could be bound and internalized via lipoprotein receptors on both Schwann cells and regenerating axons. We have asked whether the secreted apo-E in injured nerve is complexed with lipid, whether these apo-E complexes and apo-E-containing plasma lipoproteins can be taken up by neuronal growth cones, and whether the uptake is mediated by apo-B,E(LDL) receptors.

Conditioned medium containing apo-E was obtained from cultures of injured segments of rat sciatic nerves 2 weeks after crush injury, as described (4). Newly synthesized and secreted apo-E was obtained by incubating the injured segments with [35 S]methionine. To determine whether both the accumulated and newly synthesized apo-E were associated with lipid, the conditioned medium was subjected to density-gradient ultracentrifugation (Fig. 1). SDS-

M. J. Ignatius and E. M. Shooter, Department of Neurobiology, Stanford University School of Medicine, Stanford, CA 94305.

R. E. Pitas and R. W. Mahley, Gladstone Foundation Laboratories for Cardiovascular Disease, Cardiovascular Research Institute, Departments of Pathology and Medicine, University of California, San Francisco, CA 94140-0608.

ATTACHMENT 7

PewResearchCenter

U.S. SURVEY RESEARCH

Questionnaire design

Perhaps the most important part of the survey process is the creation of questions that accurately measure the opinions, experiences and behaviors of the public. Accurate random sampling and high response rates will be wasted if the information gathered is built on a shaky foundation of ambiguous or biased questions. Creating good measures involves both writing good questions and organizing them to form the questionnaire.

Questionnaire design is a multistage process that requires attention to many details at once. Designing the questionnaire is complicated because surveys can ask about topics in varying degrees of detail, questions can be asked in different ways, and questions asked earlier in a survey may influence how people respond to later questions. Researchers also are often interested in measuring change over time and therefore must be attentive to how opinions or behaviors have been measured in prior surveys.

Surveyors may conduct pilot tests or focus groups in the early stages of questionnaire development in order to better understand how people think about an issue or comprehend a question. Pretesting a survey is an essential step in the questionnaire design process to evaluate how people respond to the overall questionnaire and specific questions.

For many years, surveyors approached questionnaire design as an art, but substantial research over the past thirty years has demonstrated that there is a lot of science involved in crafting a good survey questionnaire. Here, we discuss the pitfalls and best practices of designing questionnaires.

Question development

There are several steps involved in developing a survey questionnaire. The first is identifying what topics will be covered in the survey. For Pew Research Center surveys, this involves thinking about what is happening in our nation and the world and what will be relevant to the public, policymakers and the media. We also track opinion on a variety of issues over time so we often ensure that we update these trends on a regular basis so we can understand whether people's opinions are changing.

At Pew Research Center, questionnaire development is a collaborative and iterative process where staff meet to discuss drafts of the questionnaire several times over the course of its development. After the questionnaire is drafted and reviewed, we pretest (<http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/#pretests>) every questionnaire and make final changes before fielding the survey.

Measuring change over time

Many surveyors want to track changes over time in people's attitudes, opinions and behaviors. To measure change, questions are asked at two or more points in time. A cross-sectional design, the most common one used in public opinion research, surveys different people in the same population at multiple points in time. A panel or longitudinal design, frequently used in

other types of social research, surveys the same people over time. Pew Research Center launched its own random sample panel survey in 2014; for more, see the section on the American Trends Panel (<http://www.pewresearch.org/methodology/u-s-survey-research/collecting-survey-data/#atp>).

Many of the questions in Pew Research surveys have been asked in prior polls. Asking the same questions at different points in time allows us to report on changes in the overall views of the general public (or a subset of the public, such as registered voters, men or African Americans).

When measuring change over time, it is important to use the same question wording and to be sensitive to where the question is asked in the questionnaire to maintain a similar context as when the question was asked previously (see question wording (<http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/#question-wording>) and question order (<http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/#question-order>) for further information). All of our survey reports include a topline questionnaire that provides the exact question wording and sequencing, along with results from the current poll and previous polls in which the question was asked.

Open- and closed-ended questions

One of the most significant decisions that can affect how people answer questions is whether the question is posed as an open-ended question, where respondents provide a response in their own words, or a closed-ended question, where they are asked to choose from a list of answer choices.

For example, in a poll conducted after the presidential election in 2008, people responded very differently to two versions of this question: “What one issue mattered most to you in deciding how you voted for president?” One was closed-ended and the other open-ended. In the closed-ended version, respondents were provided five options (and could volunteer an option not on the list).

When explicitly offered the economy as a response, more than half of respondents (58%) chose this answer; only 35% of those who responded to the open-ended version volunteered the economy. Moreover, among those asked the closed-ended version, fewer than one-in-ten (8%) provided a response other than the five they were read; by contrast fully 43% of those asked the open-ended version provided a response not listed in the closed-ended version of the question. All of the other issues were chosen at least slightly more often when explicitly offered in the closed-ended version than in the open-ended version. (Also see “High Marks for the Campaign, a High Bar for Obama” (<http://www.people-press.org/2008/11/13/high-marks-for-the-campaign-a-high-bar-for-obama/>) for more information.)

Researchers will sometimes conduct a pilot study using open-ended questions to discover which answers are most common. They will then develop closed-ended questions that include the most common responses as answer choices. In this way, the questions may better reflect what the public is thinking or how they view a particular issue.

Fewer People Mention Economy in Open-Ended Version

What one issue mattered most to you in deciding how you voted for president?

	Open-ended ¹	Closed-ended ²
The economy	35%	58%
The war in Iraq	5	10
Health care	4	8
Terrorism	6	8
Energy policy	*	6
Other	43	8
Candidate mentions	9	-
Moral values/social issues	7	-
Taxes/dist. of income	7	-
Other issues	5	-
Other political mentions	3	-
Change	3	-
Other	9	-
Don't know	7	2
	100	100

¹Data from Pew Research November 2008 Post-election survey

²Unprompted first response to open-ended question

* First choice from five options read to respondents

When asking closed-ended questions, the choice of options provided, how each option is described, the number of response options offered and the order in which options are read can all influence how people respond. One example of the impact of how categories are defined can be found in a Pew Research poll conducted in January 2002: When half of the sample was asked whether it was “more important for President Bush to focus on domestic policy or foreign policy,” 52% chose domestic policy while only 34% said foreign policy. When the category “foreign policy” was narrowed to a specific aspect – “the war on terrorism” – far more people chose it; only 33% chose domestic policy while 52% chose the war on terrorism.

In most circumstances, the number of answer choices should be kept to a relatively small number – just four or perhaps five at most – especially in telephone surveys. Psychological research indicates that people have a hard time keeping more than this number of choices in mind at one time. When the question is asking about an objective fact, such as the religious affiliation of the respondent, more categories can be used. For example, Pew Research Center’s standard religion question includes 12 different categories, beginning with the most common affiliations (Protestant and Catholic). Most respondents have no trouble with this question because they can just wait until they hear their religious tradition read to respond.

What is your present religion, if any? Are you Protestant, Roman Catholic, Morman, Orthodox such as Greek or Russian Orthodox, Jewish, Muslim, Buddhist, Hindu, atheist, agnostic, something else, or nothing in particular?

In addition to the number and choice of response options offered, the order of answer categories can influence how people respond to closed-ended questions. Research suggests that in telephone surveys respondents more frequently choose items heard later in a list (a “recency effect”).

Because of concerns about the effects of category order on responses to closed-ended questions, many sets of response options in Pew Research Center’s surveys are programmed to be randomized (when questions have two or more response options) to ensure that the options are not asked in the same order for each respondent. For instance, in the example discussed above about what issue mattered most in people’s vote, the order of the five issues in the closed-ended version of the question was randomized so that no one issue appeared early or late in the list for all respondents. Randomization of response items does not eliminate order effects, but it does ensure that this type of bias is spread randomly.

Questions with ordinal response categories – those with an underlying order (e.g., excellent, good, only fair, poor OR very favorable, mostly favorable, mostly unfavorable, very unfavorable) – are generally not randomized because the order of the categories conveys important information to help respondents answer the question. Generally, these types of scales should be presented in order so respondents can easily place their responses along the continuum, but the order can be reversed for some respondents. For example, in one of the Pew Research Center’s questions about abortion, half of the sample is asked whether abortion should be “legal in all cases, legal in most cases, illegal in most cases, illegal in all cases” while the other half of the sample is asked the same question with the response categories read in reverse order, starting with “illegal in all cases.” Again, reversing the order does not eliminate the recency effect but distributes it randomly across the population.

Question wording

The choice of words and phrases in a question is critical in expressing the meaning and intent of the question to the respondent and ensuring that all respondents interpret the question the same way. Even small wording differences can substantially affect the answers people provide.

An example of a wording difference that had a significant impact on responses comes from a January 2003 Pew Research Center survey. When people were asked whether they would “favor or oppose taking military action in Iraq to end Saddam Hussein’s rule,” 68% said they favored military action while 25% said they opposed military action. However, when asked whether they would “favor or oppose taking military action in Iraq to end Saddam Hussein’s rule *even if it meant that U.S. forces might suffer thousands of casualties*,” responses were dramatically different; only 43% said they favored military action, while 48% said they opposed it. The introduction of U.S. casualties altered the context of the question and influenced whether people favored or opposed military action in Iraq.

There has been a substantial amount of research to gauge the impact of different ways of asking questions and how to minimize differences in the way respondents interpret what is being asked. The issues related to question wording are more numerous than can be treated adequately in this short space. Here are a few of the important things to consider in crafting survey questions:

First, it is important to ask questions that are clear and specific and that each respondent will be able to answer. If a question is open-ended, it should be evident to respondents that they can answer in their own words and what type of response they should provide (an issue or problem, a month, number of days, etc.). Closed-ended questions should include all reasonable responses (i.e., the list of options is exhaustive) and the response categories should not overlap (i.e., response options should be mutually exclusive).

It is also important to ask only one question at a time. Questions that ask respondents to evaluate more than one concept (known as double-barreled questions) – such as “How much confidence do you have in President Obama to handle domestic and foreign policy?” – are difficult for respondents to answer and often lead to responses that are difficult to interpret. In this example, it would be more effective to ask two separate questions, one about domestic policy and another about foreign policy.

In general, questions that use simple and concrete language are more easily understood by respondents. It is especially important to consider the education level of the survey population when thinking about how easy it will be for respondents to interpret and answer a question. Double negatives (e.g., do you favor or oppose *not* allowing gays and lesbians to legally marry) or unfamiliar abbreviations or jargon (e.g., ANWR instead of Arctic National Wildlife Refuge) can result in respondent confusion and should be avoided.

Similarly, it is important to consider whether certain words may be viewed as biased or potentially offensive to some respondents, as well as the emotional reaction that some words may provoke. For example, in a 2005 Pew Research survey, 51% of respondents said they favored “making it legal for doctors to give terminally ill patients the means to end their lives,” but only 44% said they favored “making it legal for doctors to assist terminally ill patients in committing suicide.” Although both versions of the question are asking about the same thing, the reaction of respondents was different. In another example, respondents have reacted differently to questions using the word “welfare” as opposed to the more generic “assistance to the poor.” Several experiments have shown that there is much greater public support for expanding “assistance to the poor” than for expanding “welfare.”

One of the most common formats used in survey questions is the “agree-disagree” format. In this type of question, respondents are asked whether they agree or disagree with a particular statement. Research has shown that, compared with the better educated and better informed, less educated and less informed respondents have a greater tendency to agree with such

statements. This is sometimes called an “acquiescence bias” (since some kinds of respondents are more likely to acquiesce to the assertion than are others). A better practice is to offer respondents a choice between alternative statements. A Pew Research Center experiment with one of its routinely asked values questions illustrates the difference that question format can make. Not only does the forced choice format yield a very different result overall from the agree-disagree format, but the pattern of answers among better- and lesser-educated respondents also tends to be very different.

Acquiescence Bias

Agree-Disagree Format

*The best way to ensure peace is through military strength
(55% agree, 42% disagree)*

Forced Choice Format

The best way to ensure peace is through military strength (33%)

OR

Diplomacy is the best way to ensure peace (55%)

PEW RESEARCH CENTER Agree-Disagree question from Oct. 1999. Forced choice question from Sep. 1999.

One other challenge in developing questionnaires is what is called “social desirability bias.” People have a natural tendency to want to be accepted and liked, and this may lead people to provide inaccurate answers to questions that deal with sensitive subjects. Research has shown that respondents understate alcohol and drug use, tax evasion and racial bias; they also may overstate church attendance, charitable contributions and the likelihood that they will vote in an election. Researchers attempt to account for this potential bias in crafting questions about these topics. For instance, when Pew Research Center surveys ask about past voting behavior, it is important to note that circumstances may have prevented the respondent from voting: “In the 2012 presidential election between Barack Obama and Mitt Romney, did things come up that kept you from voting, or did you happen to vote?” The choice of response options can also make it easier for people to be honest; for example, a question about church attendance might include three of six response options that indicate infrequent attendance. Research has also shown that social desirability bias can be greater when an interviewer is present (e.g., telephone and face-to-face surveys) than when respondents complete the survey themselves (e.g., paper and web surveys).

Lastly, because slight modifications in question wording can affect responses, identical question wording should be used when the intention is to compare results to those from earlier surveys (see measuring change over time (<http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/#measuring-change-over-time>) for more information). Similarly, because question wording and responses can vary based on the mode used to survey respondents, researchers should carefully evaluate the likely effects on trend measurements if a different survey mode will be used to assess change in opinion over time (see collecting survey data (<http://www.pewresearch.org/methodology/u-s-survey-research/collecting-survey-data/>) for more information).

Question order

Once the survey questions are developed, particular attention should be paid to how they are ordered in the questionnaire. The placement of a question can have a greater impact on the result than the particular choice of words used in the question.

When determining the order of questions within the questionnaire, surveyors must be attentive to how questions early in a questionnaire may have unintended effects on how respondents answer subsequent questions. Researchers have demonstrated that the order in which questions are asked can influence how people respond; earlier questions – in particular those directly preceding other questions – can provide context for the questions that follow (these effects are called “order effects”).

One kind of order effect can be seen in responses to open-ended questions. Pew Research surveys generally ask open-ended questions about national problems, opinions about leaders and similar topics near the beginning of the questionnaire. If closed-ended questions that relate to the topic are placed before the open-ended question, respondents are much more likely to mention concepts or considerations raised in those earlier questions when responding to the open-ended question.

For closed-ended opinion questions, there are two main types of order effects: contrast effects, where the order results in greater differences in responses, and assimilation effects, where responses are more similar as a result of their order.

More People Favor Civil Unions When Asked After Gay Marriage

Asked first	<i>Legal agreements</i>	%	<i>Gay marriage</i>	%
	Favor	37	Favor	33
	Oppose	55	Oppose	61
	Don't know	8	Don't know	6
		100		100
Asked second	<i>Gay marriage</i>		<i>Legal agreements</i>	
	Favor	30	Favor	45
	Oppose	58	Oppose	47
	Don't know	12	Don't know	8
		100		100
N		780		735

PEW RESEARCH CENTER Oct. 2003.

An example of a contrast effect can be seen in a Pew Research Center poll conducted in October 2003 that found that people were more likely to favor allowing gays and lesbians to enter into legal agreements that give them the same rights as married couples when this question was asked after one about whether they favored or opposed allowing gays and lesbians to marry (45% favored legal agreements when asked after the marriage question, but 37% favored legal agreements without the immediate preceding context of a question about gay marriage). Responses to the question about gay marriage, meanwhile, were not significantly affected by its placement before or after the legal agreements question.

More Overall Dissatisfaction When Asked After Bush Approval

Asked first	<i>Overall satisfaction</i>	%	<i>Bush approval</i>	%
	Satisfied	17	Approve	25
	Dissatisfied	78	Disapprove	67
	Don't know	5	Don't know	8
		100		100
Asked second	<i>Bush approval</i>		<i>Overall satisfaction</i>	
	Approve	24	Satisfied	9
	Disapprove	68	Dissatisfied	88
	Don't know	8	Don't know	3
		100		100
N		766		723

PEW RESEARCH CENTER Dec. 2008.

Another experiment embedded in a December 2008 Pew Research poll also resulted in a contrast effect. When people were asked “All in all, are you satisfied or dissatisfied with the way things are going in this country today?” immediately after having been asked “Do you approve or disapprove of the way George W. Bush is handling his job as president?”; 88% said they were dissatisfied, compared with only 78% without the context of the prior question. Responses to presidential approval remained relatively unchanged whether national satisfaction was asked before or after it. A similar finding occurred in December 2004 when both satisfaction and presidential approval were much higher (57% were dissatisfied when Bush approval was asked first vs. 51% when general satisfaction was asked first).

Several studies also have shown that asking a more specific question before a more general question (e.g., asking about happiness with one’s marriage before asking about one’s overall happiness) can result in a contrast effect. Although some exceptions have been found, people tend to avoid redundancy by excluding the more specific question from the general rating.

More Endorse Working Together When Asked Second

Asked first	<i>Should Rep. leaders...</i>	<i>%</i>	<i>Should Dem. leaders...</i>	<i>%</i>
	Work with Obama	66	Work with Rep. leaders	82
	Stand up to Obama	28	Stand up to Rep. leaders	13
	Don't know	<u>6</u>	Don't know	<u>5</u>
		100		100
Asked second	<i>Should Dem. leaders...</i>	<i>%</i>	<i>Should Rep. leaders...</i>	<i>%</i>
	Work with Rep. leaders	71	Work with Obama	81
	Stand up to Rep. leaders	21	Stand up to Obama	15
	Don't know	<u>8</u>	Don't know	<u>4</u>
		100		100
N		744		756

PEW RESEARCH CENTER Nov. 2008 Post-election survey.

Assimilation effects occur when responses to two questions are more consistent or closer together because of their placement in the questionnaire. We found an example of an assimilation effect in a Pew Research poll conducted in November 2008 when we asked whether Republican leaders should work with Obama or stand up to him on important issues and whether Democratic leaders should work with Republican leaders or stand up to them on important issues. People are more likely to say that Republican leaders should work with Obama when the question was preceded by the one asking what Democratic leaders should do in working with Republican leaders (81% vs. 66%). However, when people were first asked about Republican leaders working with Obama, fewer said that Democratic leaders should work with Republican leaders (71% vs. 82%).

The order questions are asked is of particular importance when tracking trends over time. As a result, care should be taken to ensure that the context is similar each time a question is asked. Modifying the context of the question could call into question any observed changes over time (see measuring change over time (<http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/#measuring-change-over-time>) for more information).

A questionnaire, like a conversation, should be grouped by topic and unfold in a logical order. It is often helpful to begin the survey with simple questions that respondents will find interesting and engaging to help establish rapport and motivate them to continue to participate in the survey. Throughout the survey, an effort should be made to keep the survey interesting and not overburden respondents with several difficult questions right after one another. Demographic questions such as income,

education or age should not be asked near the beginning of a survey unless they are needed to determine eligibility for the survey or for routing respondents through particular sections of the questionnaire. Even then, it is best to precede such items with more interesting and engaging questions.

Pilot tests and focus groups

Similar to pretests (<http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/#pretests>), pilot tests are used to evaluate how a sample of people from the survey population respond to the questionnaire. For a pilot test, surveyors typically contact a large number of people so that potential differences within and across groups in the population can be analyzed. In addition, pilot tests for many surveys test the full implementation procedures (e.g., contact letters, incentives, callbacks, etc.). Pilot tests are usually conducted well in advance of when the survey will be fielded so that more substantial changes to the questionnaire or procedures can be made. Pilot tests are particularly helpful when surveyors are testing new questions or making substantial changes to a questionnaire, testing new procedures or different ways of implementing the survey, and for large-scale surveys, such as the U.S. Census.

Focus groups are very different from pilot tests because people discuss the survey topic or respond to specific questions in a group setting, often face to face (though online focus groups are sometimes used). When conducting focus groups, the surveyor typically gathers a group of people and asks them questions, both as a group and individually. Focus group moderators may ask specific survey questions, but often focus group questions are less specific and allow participants to provide longer answers and discuss a topic with others. Focus groups can be particularly helpful in gathering information before developing a survey questionnaire to see what topics are salient to members of the population, how people understand a topic area and how people interpret questions (in particular, how framing a topic or question in different ways might affect responses). For these types of focus groups, the moderator typically asks broad questions to help elicit unedited reactions from the group members, and then may ask more specific follow-up questions.

For some projects, focus groups may be used in combination with a survey questionnaire to provide an opportunity for people to discuss topics in more detail or depth than is possible in the interview. An important aspect of focus groups is the interaction among participants. While focus groups can be a valuable component of the research process, providing a qualitative understanding of the topics that are quantified in survey research, the results of focus groups must be interpreted with caution. Because people respond in a group setting their answers can be influenced by the opinions expressed by others in the group, and because the total number of participants is often small (and not a randomly selected subset of the population), the results from focus groups should not be used to generalize to a broader population.

Pretests

One of the most important ways to determine whether respondents are interpreting questions as intended and whether the order of questions may influence responses is to conduct a pretest using a small sample of people from the survey population. The pretest is conducted using the same protocol and setting as the survey and is typically conducted once the questionnaire and procedures have been finalized.

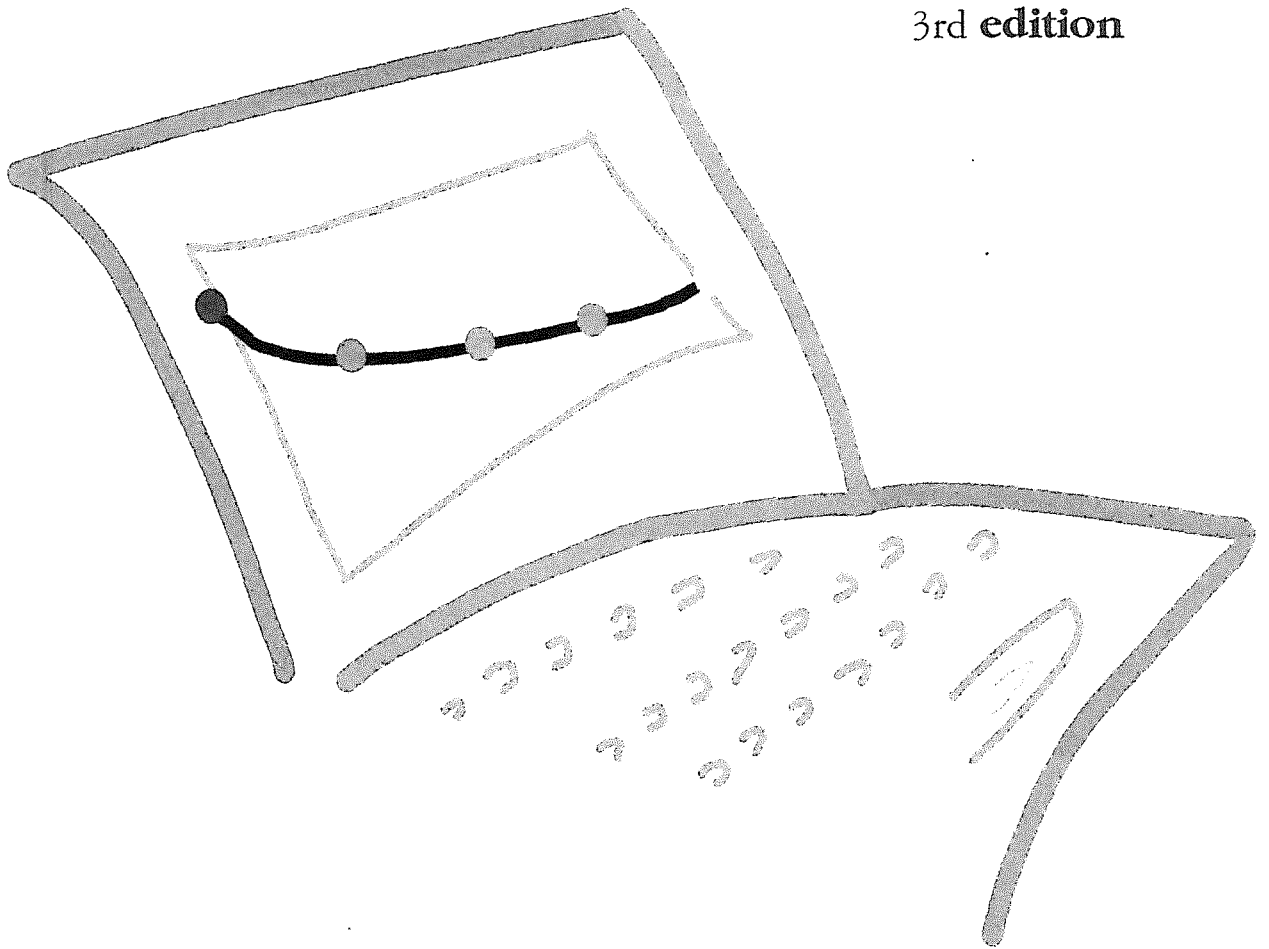
For telephone surveys, interviewers call respondents as they would in the actual survey. Surveyors often listen to respondents as they complete the questionnaire to understand if there are problems with particular questions or with the order questions are asked. In addition, surveyors get feedback from interviewers about the questions and an estimate of how much time it will take people to respond to the questionnaire.

Pew Research Center pretests all of its questionnaires, typically on the evening before a survey is scheduled to begin. The staff then meet the following day to discuss the pretest and make any changes to the questionnaire before the survey goes into the field. Information from pretesting is invaluable when making final decisions about the survey questionnaire.

ATTACHMENT 8

the **survey** research
h a n d b o o k

3rd **edition**



pamela l. **alreck** robert b. **settle**

For Expressing Questions Correctly

Guidelist 4-1

1. Use only core vocabulary—the words and phrases people use in casual speech.
2. Limit the vocabulary so the least sophisticated respondent would be familiar with the words.
3. Use simple sentences where possible, and complex sentences only when they're actually required.
4. Use two or more short, simple sentences rather than one compound or compound-complex sentence.
5. Change long, dependent clauses in sentences to words or short phrases where possible.

Instrumentation Bias and Error

The way questions are expressed can all too often introduce systematic bias, random error, or both. Even questions expressed with focus, brevity, and clarity may jeopardize reliability and/or validity. Use of the proper vocabulary and grammar doesn't guarantee that they'll be free from bias or error. Consequently, several kinds of instrumentation bias and error and the means of avoiding them must be noted.

Unstated Criteria

If the criteria by which respondents must judge some issue or respond to some question aren't completely obvious, the criteria must be stated in the question. If an item might be judged by multiple standards and the criteria aren't explicitly stated, some respondents will use one set of criteria and others will use another.

In Example 4-6, there's no clear indication of the criterion in the incorrect question. Thus, some people may respond based on their own needs and others may consider what the stores need to do to win customers in general. The correct question clearly indicates the criterion to be the personal preference of the respondent only.

The Use of Unstated Criteria

Example 4-6

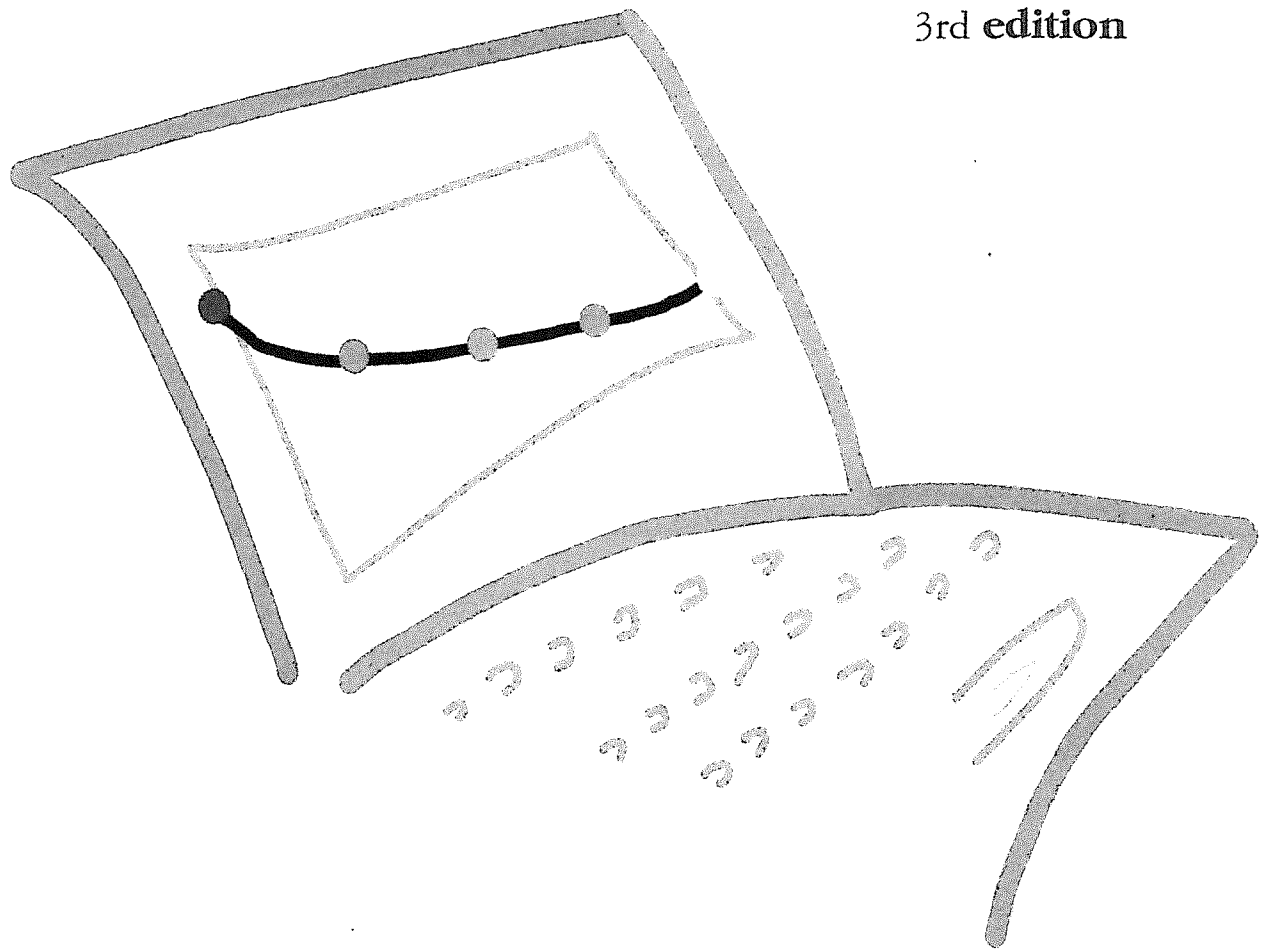
Wrong: How important is it for stores to carry a large variety of different brands of this product?

Right: How important is it to you that the store you shop at carries a large variety of different brands?

ATTACHMENT 9

the **survey** research
h a n d b o o k

3rd edition



pamela l. alreck robert b. **settle**

60-second television commercial, a 30-second television commercial, or no television commercial on the subjects' images of a brand of premium ice cream. An experiment is arranged to show the 60-second commercial to a group of men at a hotel conference room on a Saturday afternoon and get their image ratings. A group of women agree to view the 30-second spot and rate their images in a school classroom before a Wednesday evening PTA meeting. Lastly, a few weeks later, a group of elderly, retired men and women at a senior center who did not see either commercial serve as a "control group" by rating the ice cream brand on a Monday morning.

If there are differences in image ratings among the three groups of subjects, to what might those differences be attributed? Perhaps the 30-second and 60-second commercials did, in fact, *cause* the image ratings to differ. Perhaps. But is there anything else that might have influenced the ratings? Obviously there are several things:

- Differences in the average age of the groups may have had an effect on the subjects' image of ice cream in general or this brand in particular.
- Mothers of school children may all be more favorably disposed to ice cream than were those in the other two groups.
- People who rate their images of ice cream (or any other food) at one time of the day may rate it very differently at a different time.
- Those who participate at a hotel meeting room may at a senior center or at a school classroom provide different ratings if at a different location.
- The employment and income status of subjects may influence the way they view a premium (versus inexpensive) brand of ice cream.
- Those who take part in such an experiment at one time of the week or month may have different reactions on another day of the week or time of the month.

Anything other than the experimental treatment(s) that might cause or influence the post-measure or experimental results is called a "*confound*" because it's "mixed up" with the treatment. Thus, the alternative explanations for the results of the experiment listed above are potential *confounds*. Seven such threats to the internal validity of experiments are listed in Figure B-3. They threaten *internal* validity in the sense that they jeopardize the results *within* the experiment, itself, quite aside from generalization of the results to the population at large. Some experimental designs are able to control all or nearly all of these threats, while other seriously flawed designs are vulnerable to most or all of these threats.

Seriously Flawed Designs

Two pseudoexperimental designs are shown in Figure B-4. They're referred to as *pseudoexperimental* because they only appear to be experiments. They're so seriously flawed they don't qualify as genuine experiments. They provide little or no control of the threats to internal validity.

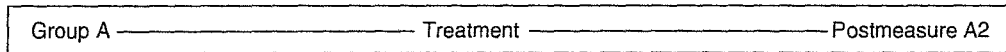
With one-group designs, there is a treatment group only and no control or comparison groups. In the top section of Figure B-4 there is only a treatment and a postmeasurement. The treatment is *assumed* to have caused or affected the

FIGURE B-3
Threats
to Internal
Validity

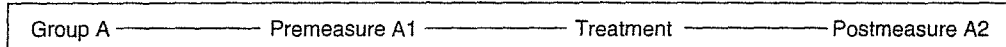
Source of Bias	Means of Control
External Events over Time	
During the experiment, an external event other than the treatment may cause a difference between the premeasurement and postmeasurement.	Avoid the seriously flawed, "Experimental Group Only" design in favor of a genuine experimental and control group design.
Internal Processes over Time	
Changes within subjects during the experiment, aside from external events, may cause a difference between pre- and postmeasurement.	Task and measure both experimental and control groups in the same way, at the same time, and over the same time period.
Premeasurement Sensitization	
The premeasurement may cause subjects to be more or less sensitive to the treatment than they would be if there were no premeasure.	Randomly assign subjects to treatments and use a postmeasure only design or the four-group, six-measure design.
Measurement Instability	
The measurement instrument or person doing the measurement may change in the interval between premeasurement and postmeasurement.	Use the same, standardized measurement instruments and procedures and the same human observers over treatment and control groups.
Systematic Selection	
Subjects aren't randomly assigned to the treatment or control group, so prior to the treatment, there are systematic differences between groups.	Randomly assign subjects to treatment groups and when that's impossible, match or assign subjects to treatments as closely to random as possible.
Experimental Attrition	
Subjects assigned to some treatment groups withdraw at a higher rate than those in others or in the control group.	Strive to make the treatment and control groups' tasks and conditions equally difficult or demanding.
Regression to the Mean	
Some subjects have extreme premeasure scores merely by chance, so when they're tested again, the second measure will, on average, be less extreme.	Include the whole groups on pre- and postmeasures, rather than postmeasuring only those with extreme premeasure scores.

results—the postmeasurement. Suppose, for example, a trainer prepared a set of instructions for employees who are in training, on how to operate a particular piece of equipment. To ascertain if they understand the instructions, the trainer has them read the instructions, then watches to see if they use the equipment correctly. If they do, the instructor *assumes* the instructions have been clear and helpful. But

FIGURE B-4 Pseudoexperimental Designs

Postmeasure Only, Experimental Group Only

The treatment is assumed to have created the change.

Pre- and Postmeasure, Experimental Group Only

Difference between A1 and A2 is assumed to be caused by the treatment.

that's a shaky assumption at best. There's no way to tell if they would have used the device correctly whether or not they read the instructions.

In the bottom section of Figure B-4 there is, once again, only one group, but this time there are both pre- and postmeasures. While superior to the other one-group design, this one still has serious flaws. Suppose a group of new enrollees in an adult education class are given a proficiency test (the premeasure) at the beginning of the course. On completion, the test is administered once again to determine if the scores have improved. At first glance, this might appear to be an acceptable design, but look more carefully!

Five of the seven threats to internal validity listed in Figure B-3 are uncontrolled: An external event, such as a newspaper or television series on the subject of the course may have affected what the students learned. Internal process over time, such as the experience they gained outside the classroom, may have affected their performance on the postmeasure. They may have been sensitized by the initial test, so they responded differently to the course than if they hadn't been tested in the beginning. The testing might have been different the second time, so there was measurement instability. And some of the "poor" students or those least motivated may have dropped the course, leading to experimental attrition. Thus, this design leaves a very great deal to be desired.

Genuine Experimental Designs

Designs that *randomly* assign subjects to *treatment* and *control groups* are the most pure forms of experimentation. They provide the most complete information and greatest control over threats to internal validity.

Randomized Assignment Designs

The genuine experimental designs shown in Figure B-5 provide the highest degree of control over the threats to internal validity listed in Figure B-3. In the diagrams of experimental designs, the time line runs from left to right. Thus, measurements that are directly above or below one another take place at the same point in time. As a consequence, threats to validity such as *external events* and *internal processes* over time, as well as *measurement instability* and *experimental attrition* are controlled

ATTACHMENT 10

The background of the page features a large, faint scatter plot. It consists of two distinct groups of data points: one group marked with 'x' symbols and another with 'o' symbols. Two parallel regression lines are drawn through the data, showing a positive correlation. The plot is divided into four quadrants by a vertical line and a horizontal line.

Experimental and Quasi-Experimental Designs

for Generalized Causal Inference

Shadish | Cook | Campbell

Testing

Sometimes taking a test once will influence scores when the test is taken again. Practice, familiarity, or other forms of reactivity are the relevant mechanisms and could be mistaken for treatment effects. For example, weighing someone may cause the person to try to lose weight when they otherwise might not have done so, or taking a vocabulary pretest may cause someone to look up a novel word and so perform better at posttest. On the other hand, many measures are not reactive in this way. For example, a person could not change his or her height (see Webb, Campbell, Schwartz, & Sechrest, 1966, and Webb, Campbell, Schwartz, Sechrest, & Grove, 1981, for other examples). Techniques such as item response theory sometimes help reduce testing effects by allowing use of different tests that are calibrated to yield equivalent ability estimates (Lord, 1980). Sometimes testing effects can be assessed using a Solomon Four Group Design (Braver & Braver, 1988; Dukes, Ullman, & Stein, 1995; Solomon, 1949), in which some units receive a pretest and others do not, to see if the pretest causes different treatment effects. Empirical research suggests that testing effects are sufficiently prevalent to be of concern (Willson & Putnam, 1982), although less so in designs in which the interval between tests is quite large (Menard, 1991).

Instrumentation

A change in a measuring instrument can occur over time even in the absence of treatment, mimicking a treatment effect. For example, the spring on a bar press might become weaker and easier to push over time, artifactually increasing reaction times; the component stocks of the Dow Jones Industrial Average might have changed so that the new index reflects technology more than the old one; and human observers may become more experienced between pretest and posttest and so report more accurate scores at later time points. Instrumentation problems are especially prevalent in studies of child development, in which the measurement unit or scale may not have constant meaning over the age range of interest (Shonkoff & Phillips, 2000). Instrumentation differs from testing because the former involves a change in the instrument, the latter a change in the participant. Instrumentation changes are particularly important in longitudinal designs, in which the way measures are taken may change over time (see Figure 6.7 in Chapter 6) or in which the meaning of a variable may change over life stages (Menard, 1991).¹⁵ Methods for investigating these changes are discussed by Cunningham (1991) and Horn (1991). Researchers should avoid switching instruments during a study; but

15. Epidemiologists sometimes call instrumentation changes surveillance bias.

if switches are required, the researcher should retain both the old and new items (if feasible) to calibrate one against the other (Murray, 1998).

Additive and Interactive Effects of Threats to Internal Validity

Validity threats need not operate singly. Several can operate simultaneously. If they do, the net bias depends on the direction and magnitude of each individual bias plus whether they combine additively or multiplicatively (interactively). In the real world of social science practice, it is difficult to estimate the size of such net bias. We presume that inaccurate causal inferences are more likely the more numerous and powerful are the simultaneously operating validity threats and the more homogeneous their direction. For example, a **selection-maturation** additive effect may result when nonequivalent experimental groups formed at the start of treatment are also maturing at different rates over time. An illustration might be that higher achieving students are more likely to be given National Merit Scholarships and also likely to be improving their academic skills at a more rapid rate. Both initial high achievement and more rapid achievement growth serve to doubly inflate the perceived effects of National Merit Scholarships. Similarly, a **selection-history** additive effect may result if nonequivalent groups also come from different settings and each group experiences a unique local history. A **selection-instrumentation** additive effect might occur if nonequivalent groups have different means on a test with unequal intervals along its distribution, as would occur if there is a ceiling or floor effect for one group but not for another.¹⁶

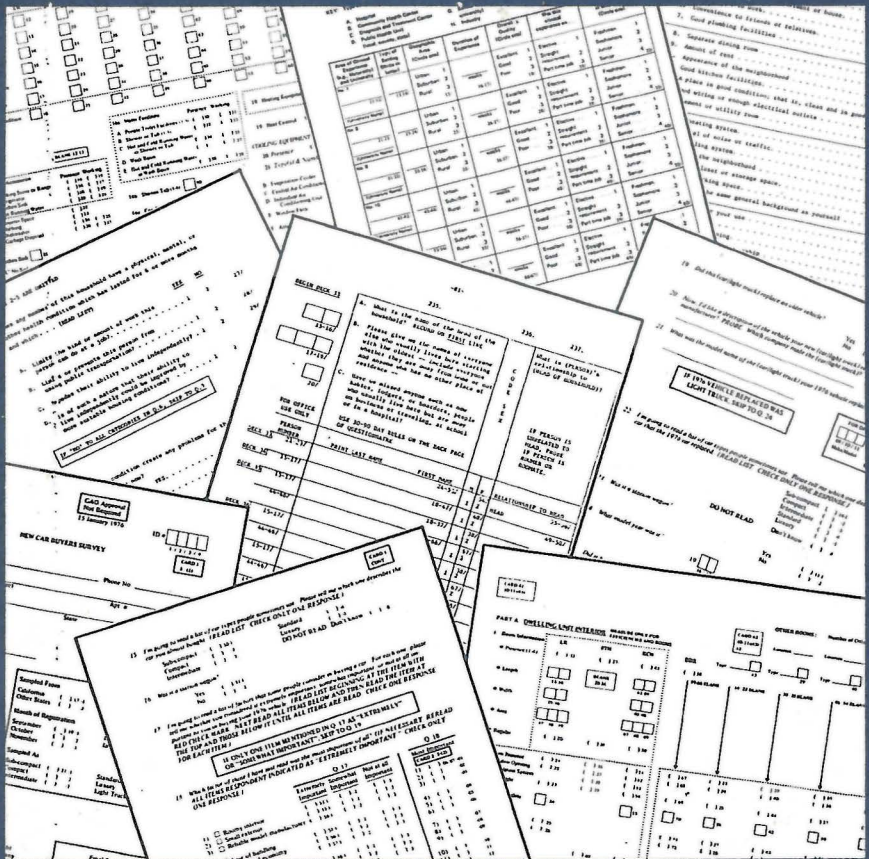
Estimating Internal Validity in Randomized Experiments and Quasi-Experiments

Random assignment eliminates selection bias definitionally, leaving a role only to chance differences. It also reduces the plausibility of other threats to internal validity. Because groups are randomly formed, any initial group differences in maturational rates, in the experience of simultaneous historical events, and in regression artifacts ought to be due to chance. And so long as the researcher administers the same tests in each condition, pretesting effects and instrumentation changes should be experienced equally over conditions within the limits of chance. So random assignment and treating groups equivalently in such matters as pretesting and instrumentation improve internal validity.

16. Cook and Campbell (1979) previously called these interactive effects; but they are more accurately described as additive. Interactions among threats are also possible, including higher order interactions, but describing examples of these accurately can be more complex than needed here.

ATTACHMENT 11

Advanced Questionnaire Design



Patricia Labaw

ings of loneliness, competence, and isolation, correlates strongly with household structures, primarily whether there are one or two heads of household. Single heads of household consistently show lower standards of living in terms of income, available transportation, and communications and social networks, and this type of household structure also correlates strongly with frequent television viewing. Many social attitudes, which correlate strongly with television viewing, in effect may result from the basic structure of the household.

□ 9

Determining respondent knowledge

The effects of respondent knowledge on answers

The problem of the effects of knowledge, or lack thereof, on respondent answers results from the fundamental operating reality of public opinion polling that people will answer questions. They will answer questions on any topic, and they will answer whether or not they know anything about that topic. The “don’t know” category supposedly takes care of people who do not know, but this assumption is mistaken. If people have even heard of a topic, they will presume to know and answer questions if encouraged by an interviewer to do so.

Given that people answer questions, how do we know whether these answers reflect people’s actual beliefs or opinions, or whether these answers simply reflect lack of knowledge? Are people really saying, “Keep the status quo as I perceive it,” or are they really saying, “I want a change”?

Knowledge is not the same as intelligence. I may not know how to perform an appendectomy, but that does not mean I am stupid. I may never have heard of automatic teller machines, but that does not mean I could not learn how to use one. I may know very little about the Palestine Liberation Organization, but that does not mean that I can’t learn anything about the political history of the Middle East.

Determining respondent lack of knowledge of specific products or public issues is crucial to understanding how to change behavior or attitudes successfully. Ascertaining levels of knowledge is the essential first step, because to change behavior, you are going to have to educate people, if possible.

From a questionnaire standpoint, people can possess three types of knowledge:

- 1) Personal knowledge derived from firsthand experience in the situation.
- 2) Factual knowledge derived from reading; personal experience; or exposure indirectly through one's profession, friends, and so forth.
- 3) Computational knowledge, which is the ability to compute or to make arithmetical calculations, including percentaging, addition, subtraction, and comparison of costs and prices.

Personal knowledge

Personal knowledge or direct experience becomes especially important in analyzing data in the areas of consumer products, corporate image, and retail services. For example, in a marketing study for the hotel/motel industry, key personal knowledge questions include: "Have you ever stayed overnight in a hotel/motel?" "How many nights per year (month or week) do you usually stay overnight in a hotel/motel?" "Do you stay while on business trips, or while on vacation trips, or while on convention trips?"

These types of knowledge questions make it possible to obtain weighted answers. Whether they consciously design for it or not, researchers must ask themselves whether *all* answers are equally important, or whether some answers are more equal than others. Just as some respondents may be weighted in a stratified probability sample, I believe that some respondents should also be weighted on the basis of the accuracy of their answers. *Knowledge, in addition to consciousness and behavior, is one of the weights that should be used to distinguish among respondents.*

The personal knowledge questions listed above are designed to sort out people who know virtually nothing from those who

know a great deal. It is a researcher and a client decision whether all answers should be treated equally, but in the event that the opinions of knowledgeable people (within the general public) should carry more weight, only identification of such people by answers to knowledge questions can allow the researcher and the client to make these important distinctions.

In short, you can sort people out through your sampling approach, and you can further sort people out by your questionnaire approach. Both are equally essential to the successful interpretation of your survey data.

Factual knowledge

Factual knowledge becomes important in the area of public policy issues. Greater knowledge may make it difficult or impossible for some people to respond to questions on such issues. People who are not knowledgeable may see the issues in a simplistic manner: "Do you favor or oppose . . .," but a knowledgeable person may ask, "Under what circumstances?" Failure to obtain a measure of this knowledge factor, as well as the ignorance factor, on public issues may lead the researcher to overstate the percentage of the public holding strong opinions on an issue, and may also obscure how quickly public opinion could change if a certain issue dominated the headlines and the facts became more widely known.

The analysis below was conducted by Michael Rappeport when he was working at Opinion Research Corporation, and the data come from two studies conducted by that organization several years ago.¹

The first poll was a regional study of attitudes toward public utilities. All respondents questioned lived in regions served by publicly owned utilities. Of particular interest to the client were answers to the question, "Which of these groups do you think *should* own and operate the electric light and power company?" The answer categories from which respondents could choose were individual investors and city, state, or federal government.

1. Michael Rappeport, "The Distinctions the Pollsters Don't Make," *The Washington Monthly*, March 1974 13-15.

The results were:

Individual investors	52%
City, state, or federal government	36
No opinion	12

Thus, only a bare majority of the general public favored private ownership of utilities, and politicians might assume that there was substantial support for public ownership.

The questionnaire for this project also contained two elementary knowledge questions:

- 1) "Which group on this card do you think owns the electric light and power company here?"
- 2) "To the best of your knowledge, does any government agency regulate the profits and rates of your local electric light and power company?"

Respondents were divided into three groups according to whether they answered both questions correctly, answered one correctly, or answered neither correctly.

When analyzed by these three nondemographic knowledge subgroups, respondents' answers to the question of which group should own and operate the local utility were distributed as follows:

"Who should own and operate the electric light and power company?"

	<i>Number of Respondents</i>	<i>Favor Investors</i>	<i>Favor Government</i>	<i>Don't Know</i>
Total respondents	1,546	52%	36	12
Answered both knowledge questions correctly	523	82%	16	2
Answered one knowledge question correctly	526	53%	37	10
Answered neither knowledge question correctly	497	12%	57	31

In other words, these respondents were telling the researcher, maintain the status quo. Those who know the utilities are owned by investors and regulated overwhelmingly prefer to maintain that situation. Those people who *think* the utilities are government

owned and not regulated prefer government ownership—the status quo, as far as they know. Neither set of answers can honestly be construed as a call for change in the ownership of local utilities.

Another example of the effects of knowledge on respondent answers occurs in the data from a poll conducted by Michael Rapoport for CBS News on problems in the Middle East.² Two factual questions were asked. One dealt with how much U.S. oil usually comes from the Arab countries, and the second dealt with whether Israel had mostly gained or mostly lost territory since it was set up in 1948. The attitude question to be analyzed by the knowledge questions was: "Some people believe the Arab countries are out to destroy Israel as a nation. Others say the Arabs are not out to destroy Israel, that they are only interested in regaining the land lost to the Israelis in the 1967 war. Which do you think is true—are the Arabs out to destroy Israel, or are they only interested in regaining their land?" Note that this attitude question presupposes knowledge of Israeli territorial gains, one of the two factual questions used for analysis.

"Are the Arabs out to destroy Israel, or are they only interested in regaining their land?"

	<i>Number of Respondents</i>	<i>Destroy Israel</i>	<i>Only Regain Land</i>	<i>Both or Neither</i>	<i>Don't Know</i>
Total respondents	1,231*	28%	52	5	15
Answered both knowledge questions correctly	351	41%	45	8	6
Answered one knowledge question correctly	328	34%	50	5	11
Answered neither knowledge question correctly	320	15%	59	4	22

*All questions were not asked of all respondents.

Notice the sharp differences in opinion between those people who knew the correct answers to both questions and those people who did not know the correct answers. Notice also that not knowing that Israel had gained land from the Arabs did not prevent people from answering "regaining the land" as the Arabs' goal.

2. *Ibid.*

These data do not say that the better informed people should be taken more seriously. On public policy issues the better informed people are often more passionate and unobjective from a policymaking point of view. Nonetheless, determining what people know and how it influences their opinions is essential before the researcher can present an objective, meaningful interpretation of the data for his client.

Computational knowledge

In these days of rising inflation, high interest rates, and higher grocery bills, people are constantly being interviewed about their financial status, attitudes toward the economy, rising prices, and unemployment. All of these interviews assume that people can make arithmetical computations. We have no data to support this assumption. In fact, we have quite a bit of data supporting the reverse assumption: that people cannot add, literally, and they cannot compute.

The following question was asked of a sample of the general public who have checking accounts. The answer categories were read to the respondents.

"If you had \$1,000 in a 5% savings account, about how much interest would you earn per year on that account?"

	<i>Respondents</i>
\$5 per year	3%
\$25 per year	8
\$50 per year	60
\$100 per year	1
\$500 per year	* (<.5%)
Don't know	27

Recall that these respondents are bank users. They all have checking accounts. They are not the lowest economic stratum of our society. Yet 39 percent of them cannot correctly guess even in this simplest of examples the amount of interest they would earn on their savings accounts. We hypothesize that this lack of computational skill accounts for the great marketing advantage that savings and loan associations have using their additional .25 percent interest rate. People simply are not able

to figure out that on a \$1,000 deposit the additional interest only \$2.50 per year, hardly enough to pay for the additional gasoline it takes to get to the savings and loan building.

The types of knowledge questions which can be included routinely in surveys and used as a scale to separate subgroups are usually very simple. The purpose is not to test people's knowledge, but to obtain some brief indication of levels of knowledge that relate to the topic being explored in the interview itself. Some routine types of knowledge questions I have often included are:

Regarding abortion:

"As far as you know, is there a movement in your state to limit abortions or make them illegal for most women?"

On energy and conservation issues:

"Has the federal government established some type of agency or department to be responsible for energy policy and practices?"

"What is the name of the federal agency which is responsible for energy policy and practices?"

On financial and bank marketing:

"What is the present rate of inflation in the U.S.?"

"What is the present interest rate paid on a *regular* pass book savings account?"

On political issues:

"What are the names of your senators? What is the name of your congressman?"

I try to design knowledge questions by thinking about the problem in the following way: if the respondent really *knew* that such and such was the case, would that probably alter his behavior or his attitude? In a jewelry marketing study if the respondent knew how much gold was in 14 karat gold versus 10 karat gold, would that make a difference in his purchasing behavior? If account users knew that differences in interest rates would result in only minor increases in their total dollars would they be more likely to favor a new type of account with different interest rate? If the respondent knew that a move-

ATTACHMENT 12

ASKING QUESTIONS

**A Practical Guide to
Questionnaire Design**

Seymour Sudman
Norman M. Bradburn



4

Questions for Measuring Knowledge

Although not as common as behavioral questions, knowledge questions have many uses in surveys. They may be used, for instance, by agencies such as the U.S. Department of Education to conduct national surveys to determine the literacy and educational achievement of adults. (Several examples of types of questions used are shown below.) The purpose of the studies is to measure the effectiveness of the educational process. Knowledge questions also are used for designing and implementing information programs or advertising campaigns. Information on the current public level of knowledge—for instance, about a subject such as cancer or a product such as a new electric car—is needed before an effective information campaign can be mounted. Measurement of the effectiveness of the information campaign requires additional surveys of information level after the campaign has been conducted.

Again, before public *attitudes* on issues and persons can be determined, it is often necessary to determine the level of public awareness and its effect on attitudes. Knowledge questions are used for this purpose. They are also used as a measure of intelligence, which may be required to help explain behavioral and attitudinal variables. Finally, they are used to obtain community or organiza-

Questions for Measuring Knowledge

89

tional information from community leaders, leaders or members of organizations, residents, or those who observed or participated in a particular event.

Checklist of Major Points

1. Before asking attitude questions about issues or persons, ask knowledge questions to screen out respondents who lack sufficient information.
2. Consider whether the level of difficulty of the questions is appropriate for the purposes of the study. For new issues simple questions may be necessary.
3. Where possible, reduce the threat of knowledge questions by asking them as opinions or using phrases such as “do you happen to know” or “can you recall, offhand.”
4. When identifying persons or organizations, avoid overestimates of knowledge by asking for additional information or including fictitious names on the list.
5. If “yes-no” questions are appropriate, ask several on the same topic, to reduce the likelihood of successful guessing.
6. For knowledge questions requiring numerical answers, use open-ended questions to avoid either giving away the answer or misleading the respondent.
7. To increase reliability when obtaining information about a geographical area, use multiple key informants or individual respondents.
8. Consider the use of pictures and other nonverbal procedures for determining knowledge.
9. When attempting to determine level of knowledge, do not use mail or other procedures that allow the respondent to look things up or to consult with others.

Examples of Knowledge Questions

Knowledge of a Public Issue: Panama Canal. In the late 1970s, before the United States ratified a new treaty with Panama, opinion surveys indicated a good deal of general awareness of the issue but much less specific knowledge. The first question in Figure 19 asks whether the respondent has heard or read about the issue.

Figure 19. Questions About Panama Canal.

1. Have you heard or read about the debate over the Panama Canal Treaties?

A great deal
Quite a lot
Not very much
Not at all
Don't know
3. As far as you know, in what year is the Panama Canal to be turned over completely to the Republic of Panama, by terms of the treaty?

2000 (1999)
Incorrect
Don't know
4. The United States secured full ownership and control of the Canal Zone by way of a treaty signed with the Republic of Panama in 1903. How much, if anything, have you heard or read about the possibility of negotiations on a new Panama Canal Treaty—a great deal, a fair amount, very little, or nothing at all?

A great deal
A fair amount
Very little
Nothing at all
Don't know

Sources: Q. 1-3, Gallup (Jan. 1978); Q. 4, Opinion Research Corporation (1975; cited in Roshco, 1978).

Gallup has used the same wording for almost every important public issue, so comparisons are possible across issues. In this instance 74 percent of respondents reported that they had heard or read about the issue; on other public issues, Gallup obtained similar reports from an average of about 70 percent of comparable samples. Thus, awareness of this issue was slightly above average. Similar results were obtained with Question 4, asked by the Opinion Research Corporation. Instead of the Gallup format, ORC asked "How much, if anything, have you heard or read . . . ?" There is no clear advantage of either format over the other.

Specific knowledge, however, as measured in Questions 2 and 3, was much lower. Only 20 percent of respondents knew that the biggest U.S. aircraft carriers did not use the Canal at all, and only 26 percent knew that the Canal was to be turned over completely to the Republic of Panama in the year 2000.

Knowledge of Persons. Name recognition is critical for political candidates during election campaigns. Also, as with public issues, opinion surveys that deal with attitudes toward public figures must first determine level of awareness. Figure 20 gives three examples of Gallup questions asking about knowledge of persons. The questions are in increasing order of difficulty. The first merely asks whether the respondent has heard something about a list—in this case a list of twenty-four political figures. In this format there is a tendency for respondents to overstate knowledge of persons, either because of name confusion or because of social desirability effects.

One way of reducing this overstatement is shown in the second question. The respondent is asked "Will you tell me who each one is or what he does?" This requires more information than the first question. Another procedure for obtaining knowledge of public figures is to show their photographs and ask the respondent for their names, as in Question 3. This is even more difficult than asking who the person is or what he does, as seen by the percent of reasonably correct answers to each of the two questions.

Health Knowledge. Figure 21 presents a series of questions about cancer. Questions 1-4 are from a study conducted by the University of Illinois Survey Research Laboratory to provide guidance for a cancer information campaign. The last four are questions that have been asked by Gallup. Note that the first question is really a knowledge question but, to reduce the threat to the respondent, is worded as an opinion question.

The Metric System. As the United States has been slowly converting to the metric system, Gallup has repeatedly asked questions that indicate a low level of knowledge by the American public about metrics. These questions are shown in Figure 22. In our judgment, these are not the most useful questions to ask to determine knowledge about the metric system. The questions all ask about the relation between the current units of measure and the new

Figure 21, Cont'd.

2. If breast cancer is found early and treated right away, how likely do you think it is that a woman will be able to do most of the things she could do before? Do you think it is . . .
 - Very likely,
 - Likely, or
 - Not very likely?
 - Don't know

3. What kinds of examinations do you know of that can be done to find breast cancer in its early stages? (*Do not read categories. Circle all that apply.*)
 - Breast self-examination (*Skip to Q. 5.*)
 - Breast examination by doctor
 - Mammography (X-ray examination)
 - Other (*Specify.*) _____
 - Don't know

4. Have you ever heard of an examination a woman can do by herself to see if there are any signs that something may be wrong with her breasts?
 - Yes
 - No

5. Do you think that cigarette smoking is or is not one of the causes of lung cancer?
 - Yes, is
 - No, is not
 - No opinion

6. Do you think cancer is curable?
 - Yes, is
 - No, is not
 - No opinion

7. Do you think cancer is contagious (catching)?
 - Yes, is
 - No, is not
 - No opinion

8. Do you happen to know any of the symptoms of cancer? What?

Sources: Q. 1-4, Survey Research Laboratory, University of Illinois (1979); Q. 5-8, Gallup (various surveys).

Figure 22. Questions About Metric System.

1. As you may know, the metric system is being introduced in this country. Do you happen to know approximately how many inches there are in a meter?

Correct	13%
Incorrect	11%
Can't say	76%

2. Do you happen to know approximately how many liters there are in a gallon?

Correct	2%
Incorrect	22%
Can't say	76%

3. One hundred kilometers are equal to how many miles?

Correct	1%
Incorrect	21%
Can't say	78%

Source: Gallup (1977).

metric units, which requires a knowledge about both systems and how they are related. Only 1 or 2 percent of the population in 1977 could relate liters to gallons and kilometers to miles. Simpler questions would probably be more appropriate at this stage. Respondents might be asked what metric units might be used to measure a person's height and weight, the contents of a soft drink container, or the distance between cities. Questions that require numerical answers are almost always more difficult for respondents than nonnumerical questions.

Questions 1 and 2 in Figure 22 are preceded by the phrase "Do you happen to know . . ." This has the effect of reducing the threat of the question and also discourages guessing. On all three questions, about three fourths of respondents chose to confess their ignorance.

Information on Products and Manufacturers. Figure 23 shows two questions (taken from Payne, 1951) about products and companies. The first provides the respondent with the name of the company and asks for the names of products that company makes. The other provides the name of the brand and asks for the name of the company. These questions might be asked in studies of attitudes

Figure 23. Questions About Products and Companies.

1. What are the brand or trade names of some of the products the (Name) company makes?
2. Will you tell me what company you think makes Frigidaire refrigerators?

Source: Payne (1951).

toward a company. These attitudes, as with attitudes on public issues, would depend on knowledge about the company.

Community Informants. In a study that we conducted at NORC of integrated neighborhoods and their characteristics, it was important to obtain information about major neighborhood institutions, such as schools and churches, as well as information on community history. Figure 24 gives examples of the kinds of questions asked of community informants. In this study four community informants—a school leader, a church leader, a community organization leader, and a leading real estate broker—were asked the same set of questions. As might be expected, they did not all give identical answers, but the mean or modal response was used to characterize the neighborhood for further analysis.

Most of the information obtained from community informants could not have been obtained in any other way. Published sources were not available or were out of date. Not all community informants were equally knowledgeable. As one might expect, the school leaders knew more about schools, the church leaders more about churches, and so on. Nevertheless, the consensus data were very useful.

Resident Information About Neighborhoods. In the same study described above, information was also obtained from a sample of neighborhood residents, not only about their personal behavior and attitudes but also about the characteristics of the neighborhood in which they lived. Two of these questions are shown in Figure 25. While residents would be expected to be generally less knowledgeable than community leaders, they are better able to report whether or not the family living next door is of the same or a different race.

The last three questions in Figure 25 are taken from another NORC study. They ask the respondent to report about the physical

Figure 24. Questions Asked of Community Informants.

1. What are the names of the public, Catholic, and private schools which children in this area attend? (*Ask A–C for each school before proceeding.*)
 - A. Who is the principal there?
(Name) _____
Don't know
 - B. What would you say is the enrollment?
(Enrollment) _____
Don't know
 - C. Is (Name) below capacity, just at capacity, slightly overcrowded, or very overcrowded?
Below capacity
At capacity
Slightly overcrowded
Very overcrowded
Don't know
2. Do both blacks and whites attend this school?
Yes (*Ask A.*)
No
Don't know
 - A. Do you happen to know the percentage of blacks in the school?
(Percent) _____
Don't know
3. Could you tell me the names of the churches and temples in the area, or nearby, which people attend? (*Probe*) Any other denominations? (*Ask A–E for each church/temple before proceeding to next one.*)
 - A. Do you happen to know the name of the minister (priest, rabbi) there?
(Name) _____
Don't know
 - B. Do both blacks and whites belong to (Name), or is this an all-white or all-black church?
Both (*Ask C and D.*)
Whites only (*Ask E.*)
Blacks only
Don't know

Asking Questions

Figure 24, Cont'd.

- C. (Hand Respondent Card 2.) What were the reactions of the members when the first black family joined?
 - Majority in favor
 - Split
 - Majority opposed
 - Majority strongly opposed
 - Don't know
- D. Approximately what is the percentage of blacks in (Name)?
 - (Percent) _____
 - Don't know
- E. (Hand Respondent Card 2.) What would be the reaction of the members if a black family were interested in joining?
 - Majority in favor
 - Split
 - Majority opposed
 - Majority strongly opposed
 - Don't know
- 4. Generally, when were the first houses (apartments) built in this neighborhood?
 - (Year) _____
 - Don't know
- 5. Were these first houses (apartments) all built and sold by the same builder, or were they built by many different people?
 - Same builder
 - Many builders (Ask A.)
 - Don't know

Source: National Opinion Research Center (1968).

Figure 25. Neighborhood Information from Residents.

- 1. As far as you know, do both white and black families live in this neighborhood?
 - Yes (If R. is black, ask A; if R. is white, go to Q. 2.)
 - No (Go to Q. 3.)
 - Don't know (Go to Q. 3.)

Questions for Measuring Knowledge

Figure 25, Cont'd.

- A. Would you say that almost all of the families living in this neighborhood are black?
 - Yes
 - No
 - Don't know
 - (Go to Q. 3.)
- 2. Are there any black families living right around here?
 - Yes (Ask A-C.)
 - No
 - Don't know
- A. About how many black families live right around here?
 - (Number) _____
- B. Do you know any of their names?
 - Yes
 - No
- C. Is there a black family living next door?
 - Yes
 - No
 - Don't know
- 3. Are there any vacant lots in this block on either side of the street?
 - Yes (Ask A.)
 - No
- A. Do any of the vacant lots have one or more of these items on them?

	Yes	No
(1) Abandoned household goods	_____	_____
(2) Broken bottles	_____	_____
(3) Trash or litter	_____	_____
(4) Remains of a partially demolished structure	_____	_____
- 4. On your block, are there any vandalized or abandoned buildings or any buildings with boarded-up windows or doors, on either side of the street?
 - Yes
 - No
- 5. Is the public street or road nearest your house or building paved?
 - Yes
 - No

Source: National Opinion Research Center (Q. 1-2, 1968; Q. 3-5, 1973).

condition of the surrounding neighborhood—litter, vandalism, and road conditions. In a face-to-face interview, the interviewer may be able to obtain some of this information by observing and recording the condition of the area. This is, of course, not possible with telephone interviewing. Even with face-to-face interviewing, the resident will have a better knowledge of the area than the interviewer, especially if the questions require more than merely brief observation.

It must be recognized, however, that residents, including community leaders, are not merely disinterested observers but have large emotional stakes in their communities. Answers to factual questions may be affected by attitudes as well as by level of knowledge. Thus, single responses about a neighborhood may not be correct. Averaging the responses from the same neighborhood increases both reliability and usefulness.

Knowledge of Occupations. Figure 26 presents a series of questions used to determine how much people know about various jobs. The primary reason for these questions is to help explain how different people rate the prestige of different occupations. Obviously, one factor involved in rating is knowledge. Note that there are five dichotomous ("yes-no") questions for each job. A respondent should be able to get about half of the answers right, simply by guessing. Thus, it is the total right answers to all ten jobs that discriminates between respondents, and not the right answers to a single question or a selected job. It is also possible to compare public familiarity with individual jobs, although this was not the primary purpose of these questions.

Media Exposure. One may sometimes wish to know how many persons are aware of a new book, magazine, movie, or television program. Figure 27 gives an example of a question asked to determine knowledge about a television program. Since awareness that *Across the Fence* is a television program is a low level of information and some respondents might guess that, the other question asks the time the program is shown. Respondents might also be asked about the content of the program, the persons appearing on it, and other details, although that was not done here.

National Assessment of Educational Progress. The most ambitious program to measure the effects of education on the United

Figure 26. Questions About Various Occupations.

1. Which of the following tools would a metal caster in a foundry be likely to use?
 - A file.
 - A cold chisel.
 - A pair of tongs.
 - A casting rod.
 - A blowtorch.
2. Which of the following things would a quality checker in a manufacturing plant be likely to do? Would he be likely to:
 - Wear a business suit?
 - Operate a cash register?
 - Write reports?
 - Supervise production line workers?
 - Examine products for defects?
3. Which of the following does a newspaper proofreader do?
 - Corrects the grammar of reporters' stories.
 - Meets the public on his job.
 - Checks the work of typesetters.
 - Rewrites newspaper stories.
 - Investigates the accuracy of rumors.
4. How many of the following things does a personnel director do?
 - Administer psychological tests.
 - Write production specifications.
 - Hire people.
 - Tell workers how to do their job.
 - Sometimes handle the complaints of workers.
5. Which of the following tools would a boilermaker be likely to use? Would he use a:
 - Jack hammer?
 - Ladder?
 - Rivet gun?
 - Crowbar?
 - Welding torch?
6. How about an optician? Does he?
 - Prescribe eyeglasses?
 - Grind lenses?
 - Test your vision?
 - Use an optical scanner?
 - Take up stock options?

Figure 26, Cont'd.

7. Which of the following would a dairy scientist be likely to use?
 - A centrifuge.
 - A Klein bottle.
 - An oscilloscope.
 - A microscope.
 - A milking stool.
8. What does a dietician do? Does he:
 - Invent new recipes?
 - Draw up menus?
 - Demonstrate cooking utensils?
 - Inspect food products?
 - Sometimes work in a hospital?
9. Which of the following things would a metal engraver be likely to need?
 - A pantograph.
 - A file.
 - A hacksaw
 - A cold chisel.
 - Acid.
10. What about a geologist? What would he be likely to use?
 - A soldering iron.
 - A rock hammer.
 - A Geiger counter.
 - A library.
 - A geodesic dome.

Source: National Opinion Research Center (1965).

States public has been the National Assessment of Educational Progress, a multimillion-dollar project of the U.S Department of Education. Figure 28 presents a series of exercises used with adults to measure knowledge in social studies, science, and writing. The standard procedure has been to pay adult participants to attempt the exercises. Standard classroom testing procedures are used, and adults are tested in their homes.

The types of questions used have varied. While mainly multiple-choice questions have been used (see Questions 2 through 11), open questions also have been asked (see Question 1, which asks for reasons why a decision was made). An especially interesting

Figure 27. Question Asked to Determine Media Knowledge.

1. I'm going to read you the name of something. Would you tell me whether you think it is the name of a book, a newspaper column, a movie, a television show, or a farmer's magazine; or perhaps you have not heard of it before? The name is *Across the Fence*.
 - Book
 - Newspaper column
 - Movie
 - Television show (*Ask A.*)
 - Farmer's magazine
 - Don't know
- A. What time is it on around here? (*Record and code.*)
 - Early morning (before 8:00)
 - Morning (8:00-12:00)
 - Afternoon (12:00-5:00)
 - Evening (5:00-10:00)
 - Late night (after 10:00)
 - Don't know

Source: National Opinion Research Center (1974).

Figure 28. Selected Questions from National Assessment of Educational Progress.

1. A major American manufacturing corporation seeks to establish a branch plant in a country that has rich natural resources but very little industry. The leaders of the nation turn down the American corporation's request.

What reasons can you give for the decision made by the leaders of the foreign nation?
2. Which one of the following is the MAJOR goal of the United Nations?
 - To fight disease
 - To maintain peace
 - To spread democracy
 - To fight the Communists
 - I don't know
3. The term "monopoly" describes the situation in which the market price of goods and services is established by which one of the following?
 - Many sellers
 - A single buyer
 - Many buyers and sellers
 - A single seller or a small group of sellers
 - I don't know

Figure 28, Cont'd.

4. Which one of the following has the power to declare an act of Congress unconstitutional?
- The Congress
 - The President
 - The United States Supreme Court
 - The United States Department of Justice
 - I don't know
5. The Supreme Court ruled that it is unconstitutional to require prayer and formal religious instruction in public schools. Which one of the following was the basis for its decision?
- The requirements violated the right to freedom of speech. There was strong pressure put on the Supreme Court by certain religious minorities.
 - Religious exercises violated the principles of the separation of church and state.
 - Every moment of the valuable school time was needed to prepare students to earn a living.
 - I don't know
6. What is needed to move cars, heat hamburgers, and light rooms?
- Conservation
 - Efficiency
 - Energy
 - Friction
 - Magnetism
 - I don't know
7. In hot climates, the advantage of buildings with white surfaces is that white surfaces effectively
- absorb light.
 - diffract light.
 - reflect light.
 - refract light.
 - transmit light.
 - I don't know
8. On the average, in human females the egg is released how many days after menstruation begins?
- 2 days
 - 9 days
 - 14 days
 - 20 days
 - 24 days
 - I don't know

Figure 28, Cont'd.

9. A fossil of an ocean fish was found in a rock outcrop on a mountain. This probably means that
- fish once lived on the mountain.
 - the relative humidity was once very high.
 - the mountain was raised up after the fish died.
 - fish used to be amphibians like toads and frogs.
 - the fossil fish was probably carried to the mountain by a great flood.
 - I don't know
10. An artificial pacemaker is an electronic device used by some patients with heart disease. What does this device simulate or replace?
- The auricles
 - The ventricles
 - The node in the right auricle
 - The heart valves between the auricles and ventricles
 - The valves that control the flow of blood into the aorta
 - I don't know
11. An object starts from rest and moves with constant acceleration. If the object has a speed of 10 meters per second after 5 seconds, the acceleration of the object is
- 1m/sec²
 - 2m/sec²
 - 5m/sec²
 - 10m/sec²
 - 50m/sec²
 - I don't know
12. (Place 12" ruler, graduated cylinder, nonporous rock, spring scales, water in jar, and string in front of respondent. Give respondent the Workbook.) In front of you are a small rock and several pieces of apparatus. You are to use whatever apparatus you find necessary to find the VOLUME of the small rock. List all procedures and record all measurements you make in the Workbook in part A. I will be making the same measurements in same way that you do. When you have determined the volume of the rock, record your answer in part B.
- (If respondent does not proceed, say "Think of some measurements you could make which would give you the volume of the rock.")
- (Indicate the equipment respondent uses.)
- Graduated cylinder and water
 - Graduated cylinder and no water

Figure 28, Cont'd.

Ruler
Spring scales
String

13. Geology is the science which studies the Earth, the rocks of which it is made up, and the changes which take place at and beneath the surface.

(Take out Handout, 2 foam rubber blocks. Pick up one of the foam rubber blocks and twist it to show respondent that it is resilient and can be deformed without harm. Place foam blocks side by side, touching each other and lined up evenly, in front of respondent.)

The foam sheets represent a layer of rock in the earth's crust. Use one or both of the foam blocks to demonstrate faulting of the earth's crust; that is, show me a fault.

(Refer to page 3 to judge respondent's demonstration.)

Correct demonstration
Incorrect demonstration
I don't know
Did not attempt demonstration

14. Below are three ads from the Help Wanted section of a newspaper. Read all three ads and choose which job you would like best if you had to apply for one of them. Then write a letter applying for that job.

OFFICE HELPER: experience in light typing and filing desirable but not necessary, must have 1 year high school math and be able to get along with people. \$2.50/hr. to start. Start now. Good working conditions. Write to ACE Company, P.O. Box 100, Columbia, Texas 94082.

SALESPERSON: some experience desirable but not necessary, must be willing to learn and be able to get along with people. \$2.50/hr. to start. Job begins now. Write to ACE Shoestore, P.O. Box 100, Columbia, Texas 94082.

APPRENTICE MECHANIC: some experience working on cars desirable but not necessary, must be willing to learn and be able to get along with people. \$2.50/hr. to start. Job begins now. Write ACE Garage, P.O. Box 100, Columbia, Texas 94082.

Source: U.S. Department of Education (1972-1974).

example is Question 14, which asks the respondent to write a letter applying for a job in response to a want ad. This question is used to provide an assessment of practical writing skills.

The science questions involve not only knowledge but the use of knowledge in problem solving. In Question 12, respondents are given a ruler, a graduated cylinder, scales, water in a jar, string, and a small nonporous rock and are asked to find the volume of the rock. Other physical apparatus are used to determine knowledge. In Question 13, respondents are handed two foam rubber blocks and are told that the blocks represent a layer of rock on the earth's crust. They are then asked to use one or both of the blocks to demonstrate a fault in the earth's crust.

These examples are included to remind the reader that, in addition to standard verbal questions and responses, other methods are available for determining level of knowledge. Both respondents and interviewers usually enjoy the variety of asking and answering questions in different ways.

Culture. In a less systematic and ambitious way than the National Assessment of Educational Progress, Gallup has asked a series of questions on literature, social science, and general knowledge. A sample of these questions is given in Figure 29. It may be seen that the public is better informed about inventions than about literature, including the Bible. Another illustration of the use of a graphic procedure is Question 4. Respondents were handed an outline map of Europe and asked to identify the countries. Similar questions have used outline maps of the United States and South America.

Measuring Intelligence. This final example is taken from a study conducted at NORC (see Sudman, 1967, p. 210) to determine the qualities that make some persons better survey research interviewers than others. Since survey interviewing is a complex task, it is reasonable to expect that success would be related to intelligence. We could simply have asked the interviewers to state their IQ, but some interviewers might not wish to do so or might not know. Therefore, we measured intelligence indirectly, by asking about grades received in school or subjects liked. In addition to these indirect measures, we used a short intelligence test, adapted from the

Wechsler Adult Intelligence Scale (WAIS) Similarities Test (see the following example).

Different people see different kinds of similarities between things. In what way do you think that these pairs of things are *alike*? *

- Lion—Tiger
- Saw—Hammer
- Hour—Week
- Circle—Triangle

This scale correlated highly with the other measures used and increased the reliability of the overall measure. Note that the introduction to the question indicates that different kinds of answers are possible. As is usually the procedure in surveys, we did not mention that the test was intended as a measure of intelligence, since this could make the respondents nervous. The scoring of the results, however, is based on norms established in standard intelligence testing. This question was included in a mail survey that the respondents filled out in their homes and mailed back. In the usual situation, knowledge questions would not be asked on a mail survey, since respondents could look up the answer or ask for help. For this question, however, there would be nothing to look up; and it is unlikely, although not impossible, that respondents consulted with others.

Techniques and Strategies for Asking Knowledge Questions

Determining Level of Knowledge. The examples suggest that knowledge questions are an important part of the process of qualifying respondent opinions and should be asked before attitude questions are asked. This order is essential if the knowledge questions are to screen out respondents who do not have sufficient information to answer detailed attitude questions. Even if all respondents answer the attitude questions, respondents will be less likely to overclaim knowledge and more likely to state that they do not know or are undecided in their attitudes if knowledge questions come first. If the attitude questions are first, respondents may feel that they are expected to know about the issue and to have an opinion. On many

* These items are not the actual items used. The actual items and the answer scoring may be found by consulting the WAIS Similarities Test.

Figure 29. General Knowledge Questions.

1 Do you happen to know who wrote *Huckleberry Finn*? *From Here to Eternity*? *A Tale of Two Cities*?

	Percent Knowing
Huckleberry Finn	40
From Here to Eternity	22
A Tale of Two Cities	7

2 The following men are inventors. Can you tell me something they invented?

	Percent Knowing
Orville and Wilbur Wright	83
Alexander Graham Bell	83
Thomas Alva Edison	67
Samuel Morse	60
Eli Whitney	58
Guglielmo Marconi	36

3 Will you tell me the names of any of the first four books of the New Testament of the Bible—that is, the first four gospels?

	Percent Knowing
	35

4 Will you please tell me the number on this map which locates each of the following countries? (*A copy of an outline map of Europe was handed to each person interviewed, with each of the countries listed below identified by number.*)

	Percent Locating Correctly	
	1947	1955
England	72	England 65
Italy	72	France 63
France	65	Spain 57
Spain	53	Poland 32
Poland	41	Austria 19
Holland	38	Yugoslavia 16
Greece	33	Rumania 11
Czechoslovakia	25	Bulgaria 10
Yugoslavia	22	None of them 23
Hungary	18	Av. no. items correct 3
Rumania	17	
Bulgaria	13	

5 Will you tell me the name of the song which is our national anthem?

	Percent Knowing
	74

Figure 29, Cont'd.

6. Can you tell me what famous people (characters), living or dead, made the following statements well known?

	Percent Knowing
Hi Yo, Silver!	71
Come up and see me sometime.	61
Old soldiers never die, they just fade away.	59
I shall return.	57
Give me liberty or give me death.	48
What's up, Doc?	40
The only thing we have to fear is fear itself.	37
Speak softly and carry a big stick.	33
With malice toward none; with charity for all.	32
There's a sucker born every minute.	27
I came, I saw, I conquered.	19
The world must be made safe for democracy.	14
I have not yet begun to fight.	14

Source: Gallup (Q. 1-2, 1957; Q. 3, 1950; Q. 4, 1947 and 1955; Q. 5, 1947; Q. 6, 1958).

public issues, it is more important to know that opinion has not yet crystallized than to force an answer.

On many issues high or low levels of knowledge can be obtained, depending on the difficulty of the questions. The easiest type of question is one that asks "Have you heard or read . . . ?" For example, a question asking "Have you heard or read about the trouble between Israel and the Arab nations in the Middle East?" received 97 percent "yes" answers in a 1973 Gallup Poll. When this same type of question was made more specific, however, asking "Have you heard or read about the recent Sinai Disengagement Pact between Egypt and Israel?" it was answered "yes" by only 59 percent of respondents.

Somewhat more difficult are dichotomous and multiple-choice questions. The questions in Figures 25 and 26, which can be answered "yes" or "no," illustrate the most common kinds of dichotomous questions. Other examples from Gallup are "Do you happen to know if the federal budget is balanced; that is, does the federal government take in as much as it spends?" and "From what

you have heard or read, do you think we produce enough oil in this country to meet our present needs or do we have to import some oil from other countries?" These questions are not strictly dichotomous, since a "don't know" answer is also possible. The "don't know" answer is more likely to be given if a phrase such as "Do you happen to know" or "As far as you know" is included at the start of the question. Questions 2-10 in Figure 28 illustrate uses of multiple-choice questions, in which the alternatives are given to the respondents. These are, of course, more difficult than dichotomous questions, since the possibility of guessing the right answer is reduced. In all these questions, the answer "I don't know" is explicitly included to reduce guessing and to indicate that "don't know" answers are expected and acceptable.

More difficult still are questions that ask for details. Question 2 in Figure 20, the questions in Figure 23, and Question 2 in Figure 20 ask respondents for minimal identification about a person or company that they have heard about. This information can include titles, reason for fame, and the state or country or product that the person or company is identified with. Answering such questions correctly indicates a higher level of knowledge than does simple name recognition.

Question 3 in Figure 20 and Question 4 in Figure 29 use pictures and an outline map to determine knowledge of persons and countries. These are more difficult than providing titles or other details about public figures. Although Question 3 (Figure 20) deals with political figures, the use of pictures may be especially appropriate in identifying television and other entertainers. Another business use is to determine public familiarity with various product package designs when the brand name is removed.

At the next level of difficulty are open qualitative questions, as shown in Figure 21 (Q. 1 and Q. 3) and Figure 28 (Q. 1) and in the WAIS Similarities Test (see "Measuring Intelligence" section). While these questions vary in difficulty among themselves, they are, on the average, more difficult than the types of questions discussed so far. These questions do not usually offer an explicit choice of a "don't know" answer, since successful guessing is unlikely. Indeed, most respondents who do not know say so rather than trying to

guess, since a bad guess may be more embarrassing than a "don't know" answer.

Most difficult of all—except for special informants, such as community informants—are numerical questions. Only a handful could answer the questions in Figure 22, dealing with the metric system. Questions asking about percentages are also difficult. Aside from very important dates, such as 1492 and 1776, most dates are not well remembered. As we shall note below, efforts to make numerical questions easier by providing multiple choices introduce additional problems.

The decision on the type of question to use will depend on the researcher's needs. Questions that are either too easy or too difficult, however, will not discriminate between respondents with different levels of knowledge. As a general rule, easier knowledge questions are most appropriate for public issues in their early stages of development; more difficult questions can be asked about long-standing issues. For example, knowledge questions about the Arab-Israeli conflict in the Middle East can be at a higher level of difficulty than questions about a new national or international crisis. Similarly, in market research, questions about long-established products can be made more difficult than questions about new products.

Some advocates of particular public policies have attempted to discredit public opinion that is in opposition to their policies by demonstrating that the public knowledge of the issues is limited. While this may sometimes be legitimate, the difficulty level of the question must also be taken into account. It is always possible to find questions so difficult that virtually no respondents can answer them correctly—especially in a survey where an instant response is required and no advance warning has been given.

Reducing Threat of Knowledge Questions. As with the threatening behavior questions discussed in the previous chapter, knowledge questions raise issues of social presentation. The respondent does not wish to appear foolish or ill informed by giving obviously incorrect answers or admitting to not knowing something that everyone else knows. Much of this threat can be reduced by an introductory phrase such as "Do you happen to know" or "Can you

recall, offhand." Explicitly mentioning "I don't know" as an answer category also reduces threat. These procedures indicate that a "don't know" answer is acceptable even if it is not the most desirable answer. The use of these threat-reducing phrases reduces the amount of guessing and increases the percentage of "don't know" answers. Conversely, if you wish respondents to guess, the phrases used above should be omitted, and respondents should be asked to give "your best guess," as in this Gallup question: "Just your best guess, what proportion of persons on welfare are 'chiselers,' that is, are collecting more than they are entitled to?"

The line between knowledge and attitude or opinion questions is often blurred. Earlier (Figure 21, Q. 1A), a knowledge question about the symptoms of breast cancer was asked in the guise of an opinion question. The question that asks respondents to guess about the proportion of welfare chiselers is really an attitude question in the guise of a knowledge question. While a few respondents may actually know the correct proportion from reading news stories, most respondents will guess, and their guess will be based on their attitudes toward welfare programs in general.

Controlling for Overstatement of Knowledge. Respondents presented with a list of persons or organizations and asked whether they have heard or read something about them may find the question mildly threatening—especially if the list is long and includes many unfamiliar names (as in Q.1, Figure 20). Indicating that one has not heard anything about all or most of the names on the list suggests that one is out of touch with current affairs. Since the answers to this question cannot be checked, there is a tendency for respondents to overclaim having heard about persons and organizations. The easiest way to control for this is to ask an additional question about who the person is or what he does (as in Q. 2, Figure 20) or what the company makes (as in Q. 1, Figure 23).

In some cases, however, such additional qualifying questions may not be appropriate. For instance, in a study of knowledge about possible candidates for political office, such as President of the United States, the current position of a person may not be relevant, and the fact that he is a possible nominee may be evident from the context of the question. Similarly, in a study of attitudes toward civil rights, respondents may be asked about a list of civil rights

leaders, and additional questions about title or affiliation may be too difficult. A solution in this case is to add the name of a "sleeper"—a person whom no one would be expected to know. As an example, in a civil rights study conducted at NORC, the name of a graduate student was added to a list of civil rights leaders. About 15 percent of all respondents reported that they had heard of this graduate student. This then indicated that several other actual civil rights leaders whose names were supposedly recognized by about 15 percent of the population were, in reality, virtually unknown. We would speculate that the lower quarter of names in Question 1 of Figure 20 were virtually unknown at the time the survey was conducted.

The same procedure may be used with companies and brands in marketing research, to determine brand name awareness. Of course, when "sleepers" are used, it is important to avoid names of known persons and to make sure that the "sleeper" brand is not actually in use at a regional level or has not been used in the past.

Using Multiple Questions. It is well known that the reliability of individuals' scores on tests and scales increases with the number of items. Similarly, more reliable measures of an individual's knowledge are obtained if multiple questions are used. Particularly with dichotomous or multiple-choice questions, single questions are subject to high unreliability because of guessing.

If knowledge is the key dependent variable, as in the National Assessment of Educational Progress, then it is evident that many questions must be asked to obtain reliable measures of knowledge. Fewer questions are needed if knowledge is to be used as an independent variable, and a single question may be sufficient if the knowledge question is to be used to screen out respondents from being asked additional questions. Note that in many of the examples given earlier—for instance, in Figure 21—multiple questions are used.

The number of questions to ask also depends on the general level of respondent information on the topic. If most respondents know nothing or very little about an issue, it will only take one or two questions to determine that.

Asking Numerical Questions. As we have already indicated, numerical questions are generally the most difficult for respondents

to answer. If given a choice of answers, most respondents will guess and choose an answer somewhere in the middle. For this reason, Payne (1951) suggested that the correct answer be put at the top or bottom of the list of alternatives. We believe an even better procedure is not to offer alternatives to the respondent but to make such questions open ended. There is no difficulty in coding such responses, since the data are numerical and can easily be processed without need for additional coding. The open question is more likely to elicit a "don't know" response than the closed question, but respondents who do volunteer an answer or a guess will be indicating knowledge or attitudes that are not distorted by the question stimulus. The Gallup metric questions in Figure 22 are open questions that use the suggested format.

Using Key Informants. The use of key informants in social science is widespread in studies of community power and influence, community decision making and innovation, collective behavior, and ecology of local institutions. Key informants can provide information that is not currently available from census data or other published sources. A key informant, however, while usually better informed than the general public, cannot be expected to know everything, and the information provided will be subject to distortion because of the attitudes or role of the informant in the community.

As an illustration, Houston and Sudman (1975) reported that, in the study discussed in the section on "Community Informants," the church informants mentioned a higher number of churches in the neighborhood than did other informants, and the community organization informants mentioned more community organizations. These unsurprising results are a function not only of the greater expertise in their areas of specialization but also of somewhat different perspectives. Thus, the church informants tended to define a neighborhood's boundaries in terms of parish boundaries or of church attendance patterns, the school informants used school boundaries, and so on.

Clearly, it is necessary to use multiple key informants to obtain reliable information about a neighborhood. These informants should be selected to represent different aspects of leadership in the community. At a minimum, we would suggest that at least three or four key informants be used for each setting and that additional

informants be added if the data are variable. The less informed the respondents, the larger will be the number required to obtain reliable information. If, instead of informants, residents are used to provide information on neighborhood ecology, a minimum sample of about ten would probably be required. While the limits of key informant data must be recognized, key informants provide data that cannot be obtained as accurately and economically by any other procedure.

Using Nonverbal Procedures. As illustrated in Figure 28 (Q. 12 and Q. 13) and Figure 29 (Q. 4), not all knowledge questions and answers must be verbal. The use of nonverbal apparatus—such as pictures, maps, music tapes, drawings, and other real-world objects—should always be considered along with standard questions. The only disadvantage to such procedures is that they may be more costly, since they require face-to-face interviewing and additional interviewer instructions and training. The advantage of using nonverbal procedures is in obtaining a more valid measure of knowledge than can be obtained from a standard question. An added advantage is that both respondents and interviewers enjoy these questions as a change of pace from standard questions.

Nonverbal procedures may be used either as stimuli or responses. Thus, in a test of music knowledge, respondents might be asked to listen to a tape of the start of Beethoven's Fifth Symphony and asked to identify the composer and composition, or they might be given the name of a composition and asked to hum a bit of it into a tape recorder. This latter procedure and other similar procedures that require recall are more difficult than the procedures that require the respondent simply to recognize the nonverbal stimulus.

Using Self-Administered Forms. As a rule, knowledge questions are not appropriate for mail surveys and other self-administered forms. In the procedures, the respondent has the chance to look up the correct answer or to consult with others. Knowledge questions can be asked on the phone as well as face-to-face since the phone conversation prevents the respondent from seeking outside help.

There are a few exceptions to this rule. The easiest knowledge question ("Have you heard or read about . . .") can be asked on a

mail survey, although not questions that are used to screen out respondents who do not know enough to have an informed opinion. Questions that appear to be asking for attitudes but are really trying to tap knowledge—for instance, the Wechsler items in the section on "Measuring Intelligence"—may also be successful in self-administered forms. Finally, for purposes of obtaining information by the use of key informants in companies or communities, self-administered forms may be superior to personal interviews. In this situation it may be desirable for the respondent to consult records and to discuss the questions with others. The resulting answers are likely to be more complete than immediate answers given in a personal interview.

Summary

Knowledge questions are used for evaluating educational achievement, for designing and implementing information programs or advertising campaigns, for determining public awareness of current issues and persons, for measuring intelligence, and for obtaining community information.

Knowledge questions vary in difficulty. The easiest questions ask whether a respondent has heard or read about a topic; the most difficult require detailed numerical information. Questions that are too easy or too difficult do not discriminate between respondents. Questions may also vary from the standard format of verbal questions by using pictures, maps, and other physical objects. Most knowledge questions are asked in personal (face-to-face or telephone) interviews, but in selected cases they may be asked in mail interviews.

Topics discussed in the chapter include procedures for reducing threat, guessing, and overclaiming knowledge; ways of asking numerical questions; and procedures for increasing reliability by using multiple knowledge questions or multiple informants.

Additional Reading

There has been little formal research on the use of knowledge questions. As may be evident from the examples in this chapter, the

Gallup organization has been one of the major users of such questions. Reference to the collections of Gallup questions (Gallup, 1972, 1978) will be useful for other examples of knowledge questions, as well as all kinds of questions.

For information on the use of the data from key informants, see *Side by Side* (Bradburn, Sudman, and Gockel, 1971). For methodological assessment of these data, see Houston and Sudman (1975).

For additional information on the use of the short intelligence test to predict survey interviewer success, see Sudman's *Reducing the Cost of Surveys* (1967, chap. 8). For detailed information on the National Assessment of Educational Progress, see U.S. Department of Education (1972-1974).

5

Measuring Attitudes: Formulating Questions

In this chapter and the subsequent one, we take up a number of topics related to attitudinal questions. A central problem for anyone trying to write about the measurement of attitudes is how to organize the discussion. Attitude measurement has so many facets, so many difficulties, that discussions of the problems tend to go off in all directions. In the absence of any clear-cut and generally accepted theory of question construction, we have somewhat arbitrarily divided our discussion into two parts. The first part, which constitutes this chapter, deals with problems of question wording. The second part, discussed in Chapter Six, deals with the ways in which questions can be answered by the respondents. The distinction between the formulation of questions and the response options is not entirely clear, as, for example, when response alternatives are built directly into the question wording. In some instances we have arbitrarily called a particular problem one of question wording or of response options.

The best advice we can offer to those starting out to write attitude questions is to plagiarize. While plagiarism is regarded as a vice in most matters, it is a virtue in questionnaire writing—assuming, of course, that you plagiarize good-quality questions. By

ATTACHMENT 13



MARKETING RESEARCH

Methodological Foundations

GILBERT A. CHURCHILL, JR.
DAWN IACOBUCCI

Ninth Edition

Individual Question Content

The researcher's previous decisions (information needed, structure and disguise, method of administration) largely control the decisions regarding individual question content. But in editing the survey, the researcher should ask some additional questions.³

IS THE QUESTION NECESSARY?

If an issue is important and it's not been adequately covered by other questions, a new question is in order. It should be framed to yield an answer with the required detail but not more than needed. For example, family consumption behavior is often explained by "stage in the life cycle," a concept captured by a composite of variables such as marital status, presence of children, and ages of the children. The presence of children indicates a dependency relationship, particularly if the youngest child is under 6. In a study using stage of life cycle as a variable, there is no need to ask the age of each child. Rather, all that is needed is one question aimed at securing the age of the youngest child, if there are children.

ARE SEVERAL QUESTIONS NEEDED INSTEAD OF ONE?

There are often situations in which several questions are needed instead of one. Consider the question, "Why do you use Crest?" One respondent may reply, "To reduce cavities." Another may reply, "Because our dentist recommends it." Obviously two different frames of reference are being employed to answer this question. The first respondent is replying in terms of current usage, whereas the second is replying in terms of initial brand choice. It would be better to break this one question down into separate questions that reflect the possible frames of reference that could be used:

- How did you first happen to use Crest?
- What is your primary reason for using it?

DO RESPONDENTS HAVE THE NECESSARY INFORMATION?

Each item should be carefully examined to see whether the typical respondent is likely to have the information sought. Respondents will give answers, but whether the answers mean anything is another matter. Sometimes we just don't want to be embarrassed to admit we don't know something in a public opinion survey, and sometimes the question is so plausible and the interviewer so credible that we assume the question has validity. For example, you could survey people about the Consumers' Rights Act on Privacy of Internet Information, say, and, 50% or more people will report a familiarity with the Act, when in fact such an act does not exist (but it sounds plausible that it might, doesn't it?). If you ask people whether they've tried SmileBrite toothpaste, some will say they have, when the brand doesn't exist. If you ask "When is the last time you saw the Jolly Green Giant (or any of a number of brand icons) on TV?" they'll say "within the past year" when in fact the tall fellow hasn't appeared in years.

From a different perspective, consider the question, "How much does your family spend on groceries in a typical week?" Unless the respondent does the grocery shopping, he or she is unlikely to know. In a situation like this, it might be helpful to

³ E.g., see Gordon Willis, *Cognitive Interviewing and Questionnaire Design* (Sage, 2003); Charles Briggs, *Learning How to Ask* (Cambridge University Press, 2002).

begin with "filter questions" to determine if the individual is indeed likely to know, e.g., "Who does the grocery shopping in your family?"⁴

Your respondents need to have the information sought, and they need to remember it. Our ability to remember various events is influenced partly by the importance of the event itself, e.g., most of us remember where we were during the attack on the World Trade Center buildings, or more pleasantly, the first car we ever owned, but many of us are unable to recall the amount of TV or the particular shows we watched last Wednesday evening, or the first brand of mouthwash we ever used, when we switched to our current brand, or why we switched. The switching information might be very important to a brand manager for mouthwashes, but it is unimportant to most individuals, a condition we have to keep in mind when designing questionnaires. We need to put ourselves in the shoes of the respondent, not those of the product manager, when deciding what information is important enough for the individual to remember.

We also need to recognize that a person's ability to remember an event is influenced by how long ago it happened. Although we might recall the television programs we watched last night, we would have more difficulty remembering what we watched last week and impossible to recall our viewing pattern of a month ago. If an event is likely to be considered relatively unimportant to most individuals, we should ask about very recent occurrences of it.

For more important events, two effects operating in opposite directions affect a respondent's ability to provide accurate answers about events that happened in some specified time period (e.g., how many times the person has seen a doctor in the last six months). Telescoping error refers to the fact that most people remember an event as having occurred more recently than it had. Recall loss means that they forget an event happened at all. The extent of the two sources of error on the accuracy of the reported information depends on the length of the reference period. For long periods, the telescoping effect is smaller whereas the recall loss effect is larger. For short periods, the reverse is true. The appropriate reference period to frame questions depends on factors such as the purchase cycle of the product category.⁵

WILL RESPONDENTS GIVE THE INFORMATION?

Even though respondents have the information, there is always a question of whether they will share it. Often respondents are flattered that they are being asked for their opinions. Participation in Nielsen television panels makes one believe that one is influencing programming choices. Rapport is quickly built in person-to-person interviewing (in the mall, on the phone), but if mail surveys and online questionnaires are at least vaguely interesting and designed well, they won't take too much of the respondents' time and most continue through the survey.

Respondents can be unable to articulate their answers, so sometimes we must design creative surveys to help them. For example, respondents might not be able to express their preferences in furniture styles, but they can certainly state which they like best when shown pictures, prototypes, hardware samples, and fabric swatches.

Sometimes the concern is that the survey question is rather sensitive, and the respondent might not wish to divulge private information. When an issue is embar-

Can the
respondent
answer?

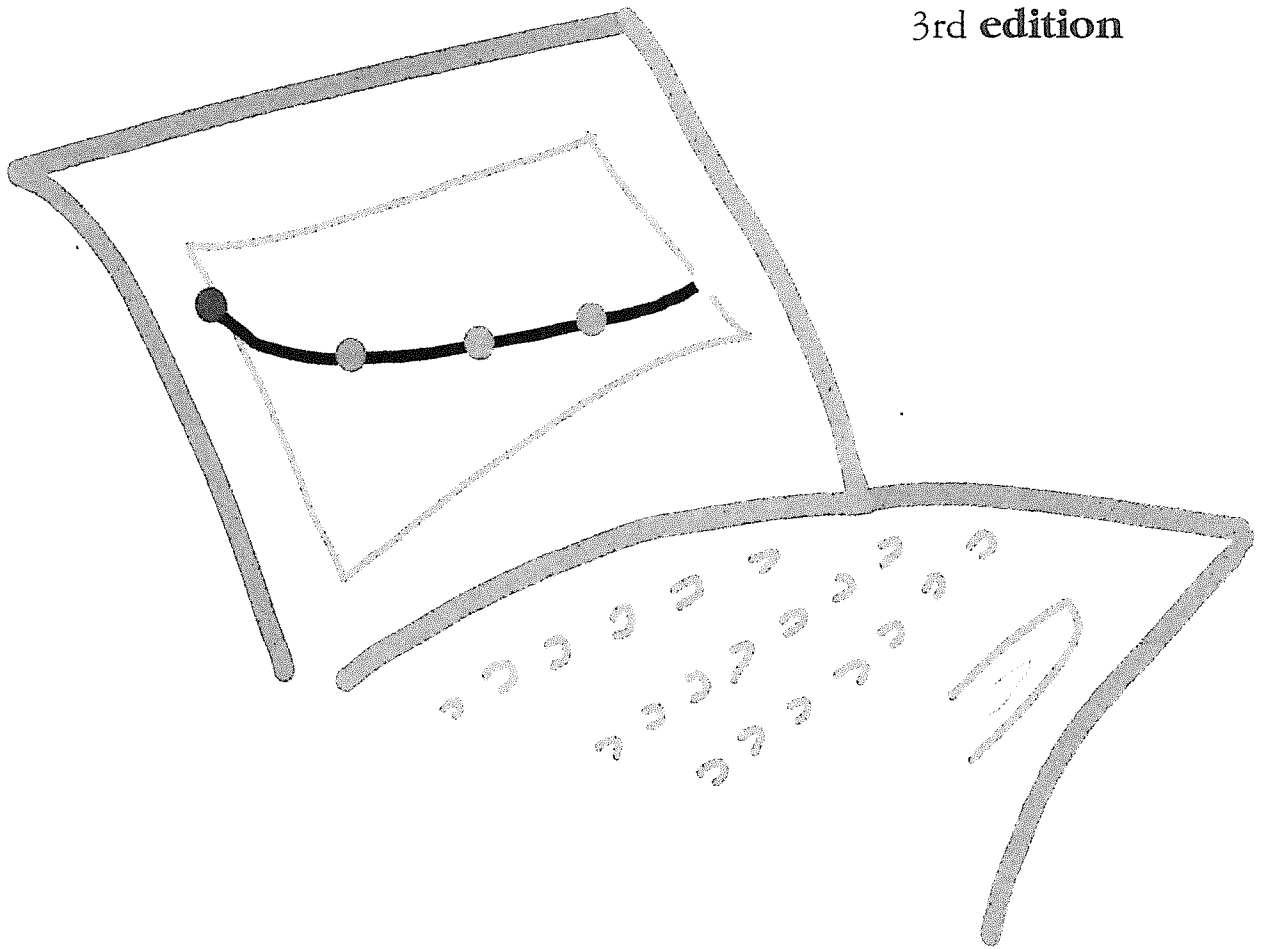
⁴ Janet Kelly and David Swindell, "The Case for the Inexperienced User: Rethinking Filter Questions in Citizen Satisfaction Surveys," *American Review of Public Administration* 33 (2003), pp. 91-108.

⁵ Other common errors include frequent users in a category underestimating their purchases, and light users overestimating their usage, cf. Eunhyu Lee, Michael Hu and Rex Toh, "Are Consumer Survey Results Distorted?" *Journal of Marketing Research* 37 (2000), pp. 125-133.

ATTACHMENT 14

the **survey** research
h a n d b o o k

3rd edition



pamela l. alreck robert b. **settle**

The Use of Overdemanding Recall

Example 4-9

Wrong: How many times did you go out on a date with your spouse before you were married?

Right: How many months were you dating your spouse before you were married?

marriage. Very few would recall that. Yet many wouldn't want to admit that they did not remember, so most respondents would probably estimate the number. Thus, there would be a large amount of error in the data. By contrast, most married respondents are likely to have good recollection of both the day when they met their spouse and the date of their marriage. These are fairly important anniversaries for most married couples. Thus, virtually all could report the number of months they were dating before their marriage.

Overgeneralizations

There are times when it may be appropriate and acceptable to ask respondents for generalizations. When a survey question seeks a generalization, it should represent a policy, strategy, or habitual pattern of the respondents, rather than specific behavior. Whenever specific incidents can be identified, the survey question should be specific. In Example 4-10, the incorrect item asks the respondents to generalize about their behavior. The question is much too general as well. It doesn't state whether the respondent is to use the immediate past as a frame of reference, go back well into the distant past, or indicate expectations for the future. By contrast, the proper wording of the item includes the past 10 incidents of the action. They would ordinarily be easily recalled by virtually all respondents, and the data are far more precise and accurate, so error is reduced and reliability is improved greatly.

Overspecificity

A survey question is overly specific when it asks for an actual or precise response that the respondent is unlikely to know or unable to express. In Example 4-11, virtually all respondents will be able to indicate their *policy* concerning the action in broad terms. Few respondents could report an exact number of times they had behaved in that way. In addition, the incorrect example would be of little value in

The Use of Overgeneralizations

Example 4-10

Wrong: When you buy "fast food," what percentage of the time do you order each of the following type of food?

Right: Of the last 10 times you bought "fast food," how many times did you eat each type of food?

The Use of Overspecificity**Example 4-11**

Wrong: When you visited the museum, how many times did you read the plaques that explain what the exhibit contained?

Right: When you visited the museum, how often did you read the plaques that explain what the exhibit contained? Would you say always, often, sometimes, rarely, or never?

many cases, because there's no *base* number of opportunities. Two people might both report they read 30 such explanations. One may have visited only 30 exhibits and read the plaque on each, while the other may have visited 300 and only read 1 in 10. They would appear from the data to be identical, yet their policies toward this issue would be very different, indeed. If the information needs were concerned with the policies, the data would lack validity.

Overemphasis

If the wording of a question is overemphatic, it's likely to introduce bias by calling for a particular type of response. When it's necessary to describe some condition in the question, it's advisable to use words that lean toward understating, rather than overstating, the condition. Respondents are then free to reach their own conclusions about the degree of severity. If the condition is described in overemphatic terms, a judgment or conclusion is imposed on the respondents. Such words as *catastrophe* or *tragedy* suggest that a potent remedy is required, while words such as *predicament* or *mishap* don't. An example of such a case is presented in Example 4-12.

The correct and incorrect examples are identical except for the last word in each: *crisis* versus *problem*. The use of the word *crisis* implies a conclusion on the part of the researcher. In addition, it's always desirable to avoid or to end a *crisis* as quickly and completely as possible. On the other hand, a *problem* isn't something that necessarily requires immediate or dramatic action. Each question must be examined carefully to avoid wording that overemphasizes or overstates the condition. Words that are overly dramatic or constitute a conclusion must be avoided.

Ambiguity of Wording

Many words and phrases designate different things for different people. Often those who write questions are totally unaware that others may have a completely

The Use of Overemphasis**Example 4-12**

Wrong: Would you favor increasing taxes to cope with the current fiscal crisis?

Right: Would you favor increasing taxes to cope with the current fiscal problem?

ATTACHMENT 15

Experiment-Research
Methodology in Marketing: Types
and Applications
Gordon L. Patzer

EXTRANEOUS VARIABLES

Hypotheses specify the effect on a dependent variable caused by an independent variable. Variables outside of these hypothesized relationships known as extraneous variables can also cause an effect on a dependent variable. Formally defined, an extraneous variable (or confounding variable) is an uncontrolled variable that causes an effect on the dependent variable. These "extra" variables confuse (or confound) data about hypothesized relationships between independent and dependent variables. If extraneous variables cause or even partially influence the data collected in an experiment, subsequent conclusions and actions will likely be erroneous. Extraneous variables in an experiment to determine the effect of container color on product sales could be other aspects of the product (appearance, quality, reputation, package size, etc.), other marketing mix variables, and environmental variables (economy, competition, laws, technology, etc.).

An extraneous variable poses a threat to experiments. The threat is that researchers are interested in the effect caused by the controlled and

manipulated variable under study, but other variables can confuse or confound the data. As a result, when conclusions are made about the effect on the dependent variable (such as product sales) caused by the independent variable (such as package color), these conclusions are likely wrong if other factors (i.e., extraneous variables) caused or even partially influenced the effect measured.

Extraneous variables in marketing research always pose threats to the accuracy of conclusions based on experiments. When an experiment is conducted in the marketplace with an aspect of the product variable as the independent variable (such as package color), extraneous variables that can impact a dependent variable (such as that product's sales) are other aspects of the product (such as quality and package size), other marketing mix variables (promotion, price, and distribution), and uncontrollable environmental variables (such as competition, economy, and societal attitudes).

The goal is to exert as much control as possible to minimize, if not eliminate, effects of these extraneous variables. A way to achieve this goal is to at least maintain the same influence of extraneous variables across all experimental conditions.

When control is not possible, which is the situation with uncontrollable environmental variables such as competition and the economy, the best procedure is to be aware of the possible influence of these extraneous variables.

As Exhibit 3.1 indicates, conclusions about cause-and-effect relationships, without careful attention to extraneous variables, are generally suspect. Therefore, one approach to controlling extraneous variables is to minimize their effects by maintaining the same influence across all experimental conditions. This control, depending on circumstances, can be fostered through either random or matched assignment of subjects. Another approach, especially with environmental variables such as competition and economy conditions, is for the researcher and user of the research to simply be aware of their possible effect, since recognition of a problem is often half the solution.

DEMAND CHARACTERISTICS

The problem of subjects responding in an experiment in a manner they think the researcher desires is just one dimension of potential demand characteristics that threaten the accuracy of an experiment. A demand charac-

teristic unintentionally provides subjects with information about the study. It consequently threatens the accuracy of data because subjects are likely to respond differently when a demand characteristic exists than when it does not. Instead of truthful responses, subjects tend to respond in a way they think researchers want or in a way they think is most socially desirable. The result ultimately leads to erroneous marketing decisions.

Demand characteristics are minimized by experiments that are properly designed and properly conducted. On the other hand, they exist when subjects know an experiment's hypothesis, independent variable(s), or other information pertinent to the research purpose. Their knowledge or awareness can come from speaking with other subjects, personal cues from a researcher such as nonverbal body language and variations in tone of voice while giving instructions, and indiscreet experimental procedures.

Demand characteristics also can occur simply by the subjects being aware that an experiment is being conducted.

Exhibit 3.1

Extraneous Variables: Reality Example

Pepsi and Michael Jackson

ATTACHMENT 16



Environmental Health - Toxic Substances

Biodegradation

Definitions

Biodegradation - "Transformation of a substance into new compounds through biochemical reactions or the actions of microorganisms such as bacteria." - U.S. Geological Survey, 2007

Biodegradation - "A process by which microbial organisms transform or alter (through metabolic or enzymatic action) the structure of chemicals introduced into the environment." - U.S. Environmental Protection Agency, 2009

Biodegradation - "Breakdown of a substance catalyzed by enzymes in vitro or in vivo. This may be characterized for purpose of hazard assessment as:

1. Primary. Alteration of the chemical structure of a substance resulting in loss of a specific property of that substance.
2. Environmentally acceptable. Biodegradation to such an extent as to remove undesirable properties of the compound. This often corresponds to primary biodegradation but it depends on the circumstances under which the products are discharged into the environment.
3. Ultimate. Complete breakdown of a compound to either fully oxidized or reduced simple molecules (such as carbon dioxide/methane, nitrate/ammonium, and water). It should be noted that the products of biodegradation can be more harmful than the substance degraded." - International Union of Pure and Applied Chemistry, 1993

Biodegradation - "[Biotransformation](#) that results in degradation of the pesticide molecule also called biodegradation, although the latter term sometimes refers to degradation processes in which the pesticide serves as a substrate for growth (e.g., Bollag and Liu, 1990)." - Nowell and others, 1999

Biodegradability (or biodegradation potential) - "The relative ease with which petroleum hydrocarbons will degrade as a result of biological metabolism. Although virtually all petroleum hydrocarbons are biodegradable, biodegradability is highly variable and dependent somewhat on the type of hydrocarbon. In general, biodegradability increases with increasing solubility; solubility is inversely proportional to molecular weight." - U.S. Environmental Protection Agency, 2009

Related Definitions

[Aerobic](#)

Toxics Home

[About The Program](#)

[GeoHealth Newsletter](#)

[Headlines](#)

[Research Projects](#)

[Contaminated Site
Management and
Remediation](#)

[Watershed-and Regional
Scale](#)

[Methods Development](#)

[Crosscutting Topics](#)

[Agricultural Chemicals](#)

[Contaminant Occurrence](#)

[Contaminant Plumes](#)

[Contaminant Transport
\(GW\)](#)

[Contaminant Transport
\(SW\)](#)

[Geophysical
Characterization](#)

[Field Methods](#)

[Laboratory Methods](#)

[Models](#)

[Natural Attenuation](#)

[Nutrients](#)

[Site Remediation](#)

[Tracer Tests](#)

[Unsaturated Zone](#)

[Publications](#)

[Search Publications](#)

[New Pubs](#)

[Photo Gallery](#)

[Frequently Asked Questions](#)

[More USGS Contaminant Info](#)

[Aerobic Biodegradation](#)

[Anaerobic](#)

[Anaerobic Biodegradation](#)

[Anoxic](#)

[Biotransformation](#)

[Electron Acceptor](#)

[Electron Donor](#)

[Natural Attenuation](#)

USGS Information on Biodegradation

- Crosscutting Topics, Toxic Substances Hydrology (Toxics) Program
 - [Natural Attenuation](#)
 - [Contaminant Plume Geochemistry and Microbiology](#)
- Toxics Program Biodegradation Investigations
 - [Biodegradation of Emerging Contaminants](#)
 - Fate of Landfill Leachate, [Norman Municipal Landfill, Norman, Oklahoma](#)
 - Sewage Contamination in Sand and Gravel Aquifers, [Cape Cod, Massachusetts](#)
 - Processes that Control the Natural Attenuation of Chlorinated Solvents
 - [Microbial Degradation of Chloroethenes in Groundwater Systems](#)
 - [Application of Molecular Methods in Microbial Ecology to Understand the Natural Attenuation of Chlorinated Solvents](#)
 - Processes Affecting the Natural Attenuation of Fuels in Groundwater
 - Crude Oil -- [Bemidji, Minnesota](#)
 - Fuel Oxygenates -- [Laurel Bay, South Carolina](#)
 - Gasoline -- [Galloway Township, New Jersey](#) [Completed]
 - Produced Water -- [Osage-Skiatook Petroleum Environmental Research Project, Oklahoma](#) [Completed]
 - [Creosote Waste in Ground Water, Pensacola, Florida](#) [Completed]
- Toxics Program Remediation Related Activities
 - [Biodegradation of Charcoal Production Wastes](#), Kingsford, Mich.
 - [Quantifying Subsurface Biodegradation](#), Norman Municipal Landfill, Norman, Okla.
 - [Can Trees Clean Up Ground Water? Phytoremediation of Trichloroethene-Contaminated Ground Water at Air Force Plant 4](#), Fort Worth, Tex.
 - [Natural Attenuation of Wood Preservatives in Ground Water](#), Pensacola, Fla.
 - [Natural Aquifer Restoration](#), Massachusetts Military Reservation, Cape Cod, Mass.
 - [RDX Biodegradation Assessment](#), Naval Submarine Base Bangor, Wash.
 - [Oxygen-Release Compound Remediation Tests](#), Laurel Bay, S.C.
 - [Quantifying Natural Attenuation at the Plume Scale](#), Galloway Township,

N.J. and Laurel Bay, S.C.

- USGS National Research Program Biodegradation Related Projects
 - [Biogeochemistry of Carbon and Nitrogen in Aquatic Environments](#)
 - [Chemical Transformations in Water Reclamation and Reuse](#)
 - [Microbial Biogeochemistry of Aquatic Environments](#)
 - [Multiphase Contaminant Transport, Reaction and Biodegradation](#)
 - [Organic Compounds in Near-Surface Environments: Understanding Fate in a Changing Biogeochemical Landscape](#)
 - [Partitioning of Solutes between Solid and Aqueous Phases](#)
 - [Subsurface Microbiology Research - Bacteria, Contaminant Interactions](#)
- [Bioremediation Activities](#), USGS Microbiology Research
- [Geochemical and Microbial Evidence of Fuel Biodegradation in a Contaminated Karst Aquifer in Southern Kentucky, June 1999](#)

Related Headlines

- [Rethinking the Limits of Oxygen-Based Biodegradation - More Oxygen Than We Think](#)
- [Complex Mixture of Contaminants Persists in Streams Miles from the Source](#)
- [Natural Attenuation Accelerates Pump-and-Treat Cleanup of TCE in Fractured Rock](#)
- [Organic Contaminants Stored on Sediments Can Slow Down Groundwater Restoration](#)
- [Sometimes the Question Is «Who Isn't Living There?»](#)
- [Do Natural Processes Mitigate Contamination from Landfill Leachate?](#)
- [Hormones Degrade in the Environment!](#)
- [Detergents in Streams May Just Disappear](#)
- [Hydrogen Measured in a New Test for Determining Subsurface Microbiological Activity at Contamination Sites](#)
- [New Report Presents a Framework for Assessing the Sustainability of Monitored Natural Attenuation](#)
- [RDX Biodegradation Under Metal-Reducing Conditions](#)
- [New Information on the Long-Term Fate of Ammonium in Ground Water](#)
- [Microorganisms Degrade MTBE Even at Winter Ground-Water Temperatures](#)
- [Does NDMA Biodegrade at Ground-Water Recharge Facilities?](#)
- [Ground-Water Recharge Affects Fate of Petroleum Contaminant Plumes](#)
- [History and Ecology of Chloroethene Biodegradation--A Review](#)
- [Using Oxygen to Enhance Biodegradation of Contaminants -- Lessons Learned](#)
- [USGS Scientists Contribute to the Landmark "Treatise on Geochemistry"](#)
- [A Unique Approach to Evaluating Natural Attenuation is Applied Worldwide](#)
- [Landmark Book Published on the Fate of Contaminants in the Environment: Partition and Adsorption of Organic Contaminants in Environmental Systems](#)
- [Using Oxygen to Clean Up Ground-Water Contamination](#)
- [MTBE Can Degrade Naturally Without Oxygen](#)
- [How Do You Clean Up Gasoline Spills Naturally?](#)
- [Relying on Nature to Clean Up Contaminated Ground Water](#)

- [MTBE Biodegrades Naturally in Stream Sediments](#)
- [Natural Attenuation for Groundwater Remediation](#)
- [Can a Sewage-Contaminated Aquifer Naturally Clean Itself?](#)
- [Natural Attenuation of MTBE at Laurel Bay, South Carolina](#)

Other Information on Biodegradation

- U.S. Environmental Protection Agency
 - [Natural Attenuation](#), CLU-IN, Office of Superfund Remediation and Technology Innovation
 - [Monitored Natural Attenuation \(MNA\)](#), Office of Research and Development
 - [Commonly Asked Questions Regarding the Use of Natural Attenuation for Petroleum Contaminated Sites at Federal Facilities](#)
 - [Commonly Asked Questions Regarding the Use of Natural Attenuation for Chlorinated Solvent Spills at Federal Facilities](#)
- National Research Council Reports
 - [Natural Attenuation for Groundwater Remediation](#)
 - [In Situ Bioremediation: When Does it Work?](#)
 - [Alternatives for Ground Water Cleanup](#)
- [Environmental Inquiry - Biodegradation](#), Cornell University and Penn State University
- [Biocatalysis/Biodegradation Database](#), University of Minnesota

References

Bollag, J.M., and Liu, S.Y., 1990, Biological transformation processes in pesticides, *in* Cheng, H.H., ed., Pesticides in the soil environment: Processes, impacts, and modeling: Madison, Wis., Soil Science Society of America, p. 169-211.

International Union of Pure and Applied Chemistry, 1993, [Glossary for chemists of terms used in toxicology](#): Pure and Applied Chemistry, v. 65, no. 9, p. 2003-2122.

Nowell, L.H., Capel, P.D., and Dileanis, P.D., 1999, Pesticides in stream sediment and aquatic biota--Distribution, trends, and governing factors: Boca Raton, Fla., Lewis Publishers, 1001 p.

U.S. Environmental Protection Agency, 2009, [Glossary of technical terms](#): U.S. Environmental Protection Agency, access date July 21, 2010.

U.S. Geological Survey, 2007, [Glossary--Biodegradation](#): U.S. Geological Survey, access date July 21, 2010.

Disclaimer: The definitions on this page are provided for information purposes only, and do not indicate endorsement by the U.S. Geological Survey.

[U.S. Department of the Interior](#) | [U.S. Geological Survey](#)

URL: <http://toxics.usgs.gov/definitions/biodegradation.html>

Page Contact Information: [Webmaster](#)

Page Last Modified: Monday, 02-Jun-2014 17:59:48 EDT

ATTACHMENT 17

edition 2
SOCIAL
RESEARCH
METHODS

Qualitative and Quantitative Approaches

H. RUSSELL BERNARD

University of Florida



Los Angeles | London | New Delhi
Singapore | Washington DC

that arithmetic skills must be emphasized during the early school years. Furthermore, it says, teachers whose classes make exceptional progress in this area should be rewarded with 10% salary bonuses.

The governor accepts the recommendation and announces a request for a special legislative appropriation. Elementary teachers all over the state start paying extra attention to arithmetic skills. Even supposing that the students in the treatment classes do better than those in the control classes, how can we be certain that the magnitude of the difference would not have been greater had this historical confound not occurred?

Maturation

The **maturation confound** refers to the fact that people in any experiment grow older, or get more experienced while you are trying to conduct an experiment. Consider the following experiment: Start with a group of teenagers on a Native American reservation and follow them for the next 60 years. Some of them will move to cities, some will go to small towns, and some will stay on the reservation. Periodically, test them on a variety of dependent variables (their political opinions, their wealth, their health, their family size, and so on). See how the various experimental treatments (city vs. reservation vs. town living) affect these variables.

Here is where the maturation confound enters the picture. The people you are studying get older. Older people in many societies become more politically conservative. They are usually wealthier than younger people. Eventually, they come to be more illness-prone than younger people. Some of the changes you measure in your dependent variables will be the result of the various treatments and some of them may just be the result of maturation.

Maturation is sometimes taken too literally. Social service delivery programs “mature” by working out bugs in their administration. People “mature” through practice with experimental

conditions and they become fatigued. We see this all the time in new social programs where people start out being enthusiastic about innovations in organizations and eventually get bored or disenchanted.

Testing and Instrumentation

The **testing confound** occurs in laboratory and field experiments when subjects get used to being tested for indicators on dependent variables. This quite naturally changes their responses. Asking people the same questions again and again in a longitudinal study, or even in an ethnographic study done over six months or more, can have this effect.

The **instrumentation confound** results from changing measurement instruments. Changing the wording of questions in a survey is essentially changing instruments. Which responses do you trust: the ones to the earlier wording or the ones to the later wording? If you do a set of observations in the field—like children’s behavior at recess or nurses’ behavior in responding to patients in a hospital or cops’ behavior in making arrests—and later send in someone else to continue the observations, you have changed instruments.

Which observations do you trust as closer to the truth: yours or those of the substitute instrument (the new field researcher)? In multi-researcher projects, this problem is usually dealt with by training all investigators to see and record things in more or less the same way. This is called increasing **interrater reliability**. (More on this in Chapter 19, on analyzing qualitative data.)

Regression to the Mean

Regression to the mean is a confound that can occur when you study groups that have extreme scores on a dependent variable. No matter what the treatment is, over time you’d expect the extreme scores to become more moderate, just because there’s nowhere else for them to go. If men who are taller than 6’7”

ATTACHMENT 18

Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews

ROBERT E. SLAVIN
Johns Hopkins University

ABSTRACT: This paper proposes an alternative to both meta-analytic and traditional reviews. The method, "best-evidence synthesis," combines the quantification of effect sizes and systematic study selection procedures of quantitative syntheses with the attention to individual studies and methodological and substantive issues typical of the best narrative reviews. Best-evidence syntheses focus on the "best evidence" in a field, the studies highest in internal and external validity, using well-specified and defended a priori inclusion criteria, and use effect size data as an adjunct to a full discussion of the literature being reviewed.

In the decade since Glass (1976) introduced the concept of meta-analysis as a means of combining results of different investigations on a related topic, the practice and theory of literature synthesis has been dramatically transformed. Scores of meta-analyses relating to educa-

Robert E. Slavin is Director, Elementary School Program, Center for Research on Elementary and Middle Schools, Johns Hopkins University, Baltimore, MD 21218. His specializations are cooperative learning, school and classroom organization, field research methods, and research review.

An earlier version of this paper was presented at the 1985 annual meeting of the American Educational Research Association, Chicago. This paper was written under grants from the National Institute of Education (No. NIE-G-83-0002) and the Office of Educational Research and Improvement (No. OERI-G-86-0006). However, the opinions expressed do not necessarily represent Department of Education policy. I would like to thank Harris Cooper, Gary Gottfredson, Nancy Madden, Robert Stevens, and Noreen Webb for their helpful comments on earlier drafts of this paper.

tional practice and policy have appeared, and the number of articles using or discussing meta-analysis in education has approximately doubled each year from 1979 to 1983 (S. Jackson, 1984). Several thoughtful guides to the proper conduct of meta-analyses have been recently published (see, e.g., Cooper, 1984; Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Light & Pillemer, 1984; Rosenthal, 1984).

Ever since it was introduced, meta-analysis has been vigorously criticized, and equally vigorously defended. In considering arguments for and against this procedure in the abstract, there is much validity to both sides. Proponents of quantitative synthesis (e.g., Cooper, 1984; Glass et al., 1981; G. Jackson, 1980; Light & Pillemer, 1984) are certainly correct to criticize traditional reviews for using unsystematic and poorly specified criteria for including studies and for using statistical significance as the only criterion of treatment effects. Critics of these procedures (e.g., Cook & Leviton, 1980; Eysenck, 1978; Slavin, 1984; Wilson & Rachman, 1983) are equal-

ly justified in objecting to a mechanistic approach to literature synthesis that sacrifices most of the information contributed in the original studies and includes studies of questionable methodological quality and questionable relevance to the issue at hand.

In an earlier article (Slavin, 1984), I evaluated the actual practice of meta-analysis in education by examining eight meta-analyses conducted by six independent sets of investigators, comparing their procedures and conclusions against the studies they included. I found that all of these meta-analyses had made errors serious enough to invalidate or call into question one or more major conclusions. In reviewing several meta-analyses published after my article went to press, I have seen misapplications of the procedure that are at least as serious (Slavin, 1985). Yet the misuses of meta-analysis in education do not in themselves justify a return to traditional review procedures.

In this paper, I propose an alternative to both meta-analytic and traditional reviews that is designed to draw on the strengths of each ap-

proach and to avoid the pitfalls characteristic of each. The main idea behind this procedure, which I call "best-evidence synthesis," is to add to the traditional scholarly literature review application of rational, systematic methods of selecting studies to be included and use of effect size (rather than statistical significance alone) as a common metric for treatment effects.

The Principle of Best Evidence

In law, there is a principle that the same evidence that would be essential in one case might be disregarded in another because in the second case there is better evidence available. For example, in a case of disputed authorship, a typed manuscript might be critical evidence if no handwritten copy is available, but if a handwritten copy exists, the typed copy would be inadmissible because it is no longer the best evidence (since the handwritten copy would be conclusive evidence of authorship).

I would propose extending the principle of best evidence to the practice of research review. For example, if a literature contains several studies high in internal and external validity, then lower quality studies might be largely excluded from the review. Let's say we have a literature with 10 randomized studies of several months' duration evaluating Treatment X. In this case, results of correlational studies, small-sample studies, and/or brief experiments might be excluded, or at most briefly mentioned. For example, Ottenbacher and Cooper (1983) located 61 randomized, double-blind studies of effects of medication on hyperactivity, and therefore decided not to include studies of lower methodological rigor. However, if a set of studies high in internal and external validity does not exist, we might cautiously examine the less well designed studies to see if there is adequate unbiased information to come to any conclusion.

The principle of best evidence works in law because there are a priori criteria for adequacy of evidence in certain types of cases. Comparable criteria could not be prescribed for all of educational research, but could be proposed for

each subfield as it is reviewed. These criteria might be derived from a reading of previous narrative and meta-analytic reviews and a preliminary search of the literature.

Justification for the "Best Evidence" Principle

The recommendation that reviewers apply consistent, well justified, and clearly stated a priori inclusion criteria is at the heart of the best-evidence synthesis, and differs from the exhaustive inclusion principle suggested by Glass et al. (1981) and others, who recommend including all studies that meet broad standards in terms of independent and dependent variables, avoiding any judgments of study quality. Proponents of meta-analysis suggest that statistical tests be used to empirically test for any effects of design features on study outcomes. The rationale given for including all studies regardless of quality rather than identifying the methodologically adequate ones is primarily that the reviewer's own biases may enter into decisions about which studies are "good" and which are "bad" methodologically. Certainly, studies of interjudge consistency in evaluations of journal articles (e.g., Gottfredson, 1978; Marsh & Ball, 1981; Peters & Ceci, 1982; Scarr & Weber, 1978) show considerable variation from reviewer to reviewer, so global decisions about methodological quality are inappropriate as a priori criteria for inclusion of studies in a research synthesis. It is important to recall that much of the impetus for the development of meta-analysis came from a frequent observation that traditional narrative reviews were unsystematic in their selection of studies, and did a poor job (or no job at all) of justifying their selection of studies, arguably the most important step in the review process (see Cooper, 1984; G. Jackson, 1980; Waxman & Walberg, 1982).

However, while it is difficult to justify a return to haphazard study selection procedures characteristic of many narrative reviews, it is also difficult to accept the meta-analysts' exhaustive inclusion strategy.

The rationale for exhaustive inclusion depends entirely on the proposition that specific methodologi-

cal features of studies can be statistically compared in terms of their effects on effect size. Cooper (1984) puts the issue this way:

If it is empirically demonstrated that studies using "good" methods produce results different from "bad" studies, the results of the good studies can be believed. When no difference is found it is sensible to retain the "bad" studies because they contain other variations in methods (like different samples and locations) that, by their inclusion, will help solve many other questions surrounding the problem area. (pp. 65-66)

In practice, meta-analyses almost always test several methodological and substantive characteristics of studies for correlations with effect size, using a criterion for rejecting the null hypothesis of no differences of .05. However, in order to justify pooling across categories of studies, the meta-analyst must prove the null hypothesis that the categories do not differ. This is logically impossible, and in situations in which the numbers of studies are small and the numbers of categories are large, finding true differences between categories of studies to be statistically significant is unlikely.

One example of this is a recent meta-analysis on adaptive education by Waxman, Wang, Anderson, and Walberg (1985), which coded the critical methodological factor "control method" into eight categories: unspecified, stratification, partial correlation, beta weights in regression, raw or metric weights in regression, factorial analysis of variance, analysis of covariance, or none. In a meta-analysis of only 38 studies, the 8×1 ANOVA apparently used to evaluate effects of methodological quality on study outcome had highly unequal and small cell sizes and an extremely high probability of failing to detect any true differences.

The problem of the reviewer's bias entering into inclusion decisions is hardly solved by exhaustive inclusion followed by statistical tests. The reviewer's bias may just as well enter into the coding of studies for statistical analysis (Mintz, 1983; Wilson & Rachman, 1983). Worse, the reader has no easy way to find out how studies were coded. For example, most of the studies coded as "randomly assigned" in a meta-

analysis on mainstreaming by Carlberg and Kavale (1980) were in fact randomly selected from non-randomly assigned groups. To discover this, it was necessary to obtain every article cited and laboriously recode them (Slavin, 1984).

Reviews of social science literature will inevitably involve judgment. No set of procedural or statistical canons can make the review process immune to the reviewer's biases. What we can do, however, is to require that reviewers make their procedures explicit and open, and we can ask that reviewers say enough about the studies they review to give readers a clear idea of what the original evidence is. The greatest problem with exhaustive inclusion is that it often produces such a long list of studies that the reviewer cannot possibly describe each one. I would argue that all other things being equal, far more information is extracted from a large literature by clearly describing the best evidence on a topic than by using limited journal space to describe statistical analyses of the entire methodologically and substantively diverse literature.

Criteria for Including Studies

Obviously, if a priori criteria are to be used to select studies, these criteria must be well thought out and well justified. It is not possible to specify in advance what criteria should be used, as this must depend on the purposes for which the review is intended (see Light & Pillemer, 1984, for more on this point). However, there are a few principles that probably apply generally.

First, the most important principle of inclusion must be germaneness to the issue at hand. For example, a meta-analysis focusing on school achievement as a dependent measure must explicitly describe what is meant by school achievement and must only include studies that measured what is commonly understood as school achievement on individual assessments, not swimming, tennis, block stacking, time-on-task, task completion rate, group productivity, attitudes, or other measures perhaps related to but not identical with student academic achievement (see Slavin, 1984).

... far more information is extracted from a large literature by clearly describing the best evidence on a topic than by using limited journal space to describe statistical analyses of the entire methodologically and substantively diverse literature.

Second, methodological adequacy of studies must be evaluated primarily on the basis of the extent to which the study design minimized bias. For example, it would probably be inappropriate to exclude studies because they failed to document the reliability of their measures, as unreliability of measures is unlikely in itself to bias a study's results in favor of the experimental or control group. On the other hand, great caution must be exercised in areas of research in which less-than-ideal research designs tend to produce systematic bias. For example, matched or correlational studies of such issues as special education, non-promotion, and gifted programs are likely to be systematically biased in favor of the students placed in regular classes, promoted, or placed in gifted classes, respectively (Madden & Slavin, 1983). In these areas of research, the independent variable is strongly correlated with academic ability, motivation, and many other factors that go into a decision to, for example, promote or retain a student.

Controlling for all these factors is virtually impossible in a correlational study. In research literatures of this kind, random assignment to experimental or control groups is essential. However, in other areas of research, the independent variable is less highly correlated with academic ability or other biasing factors. For example, schools that use tracking may not be systematically different from those that do not. If this is the case, then random assignment, though still desirable, may be less essential; carefully matched or statistically controlled studies may be interpretable.

Third, it is important to note that external validity should be valued at least as highly as internal validity in selecting studies for a best-evidence synthesis. For example, reviews of classroom practices should not generally include extremely brief laboratory studies or other highly artificial experiments. Often, a search for randomized studies turns up such artificial experiments. This was the case with the Glass, Cohen, Smith, and Filby (1982) class size meta-analysis, which found more positive effects of class size in "well controlled" studies than in "less well controlled" studies. Well controlled meant studies using random assignment, but this requirement caused the well controlled study category to include a number of extremely brief artificial experiments, such as a 30-minute study of class size by Moody, Bausell, and Jenkins (1973), as well as a study of effects of class size on tennis "achievement" (Verducci, 1969). Because class size is not strongly correlated with academic ability (see Coleman et al., 1966), this is actually a case in which well designed correlational studies, because of their greater external validity, might be preferred to many of the randomized experimental studies.

One category of studies that may be excluded in some literatures is studies with very small sample sizes. Small samples are generally susceptible to unstable effects. In education, experiments involving small numbers of classes are particularly susceptible to teacher and class effects (see Glass & Stanley, 1970; Page, 1975). For example, if Mr. Jones teaches Class A using Method X and Ms. Smith teaches Class B

using Method *Y*, there is no way to rule out the possibility that any differences between the classes are due to differences in teaching style or ability between Mr. Jones and Ms. Smith (teacher effects) or to effects of students in the different classes on one another (class effects) rather than to any differences between Methods *X* and *Y*. To minimize these possibilities, a criterion of a certain number of teachers, classes, and/or students in each treatment group might be established.

In some literatures lacking a body of studies high in internal and external validity, it may be necessary to include (but not pool) germane studies using several methods, each of which has countervailing flaws. For example, if a literature on a particular topic consists largely of randomized experiments low in external validity and correlational studies high in external validity but susceptible to bias, the two types of research might be separately reviewed. If the two groups of studies yield the same result, each buttresses the other. If they yield different results, the reviewer should explain the discrepancy.

Finally, it may be important in some literatures to mention the best designed studies excluded from the review (that is, those that "just missed") to give the reader a more concrete idea of why a study was excluded and what the consequences of that exclusion are. For example, one recent meta-analysis of studies of bilingual education by Willig (1985) devoted considerable attention to describing studies excluded from the review, making the criteria for inclusion clear.

Some arbitrary limitations often placed on inclusion of studies in traditional reviews make little sense, and should be abandoned. Perhaps most common is the elimination of dissertations and unpublished reports (such as government reports or university technical reports). Often, these unpublished reports are better designed than published ones; for example, it may sometimes be easier to get a poorly designed study into a low quality journal than to get it past a dissertation committee. The most important randomized study of special educa-

tion versus mainstream placement (Goldstein, Moss, & Jordan, 1966) and the Coleman Report (Coleman et al., 1966) are two examples of unpublished government reports essential to their respective literatures.

On the other hand, meta-analyses also exclude one type of study that should not be excluded: studies in which effect sizes cannot be computed. It often happens that studies fail to report standard deviations or other information sufficient to enable computation of effect sizes. While effect sizes can be computed directly from *t*-scores, *F*'s, or *p* values for two-group comparisons if *N*'s are known (see Glass et al., 1981), there are cases in which important, well designed studies present only *p* values or *F*'s for complex designs, ANCOVAs, or multiple regression analyses with too little information to allow for computation of effect sizes. Yet there is no good reason to exclude these studies from consideration solely on this basis.

Exhaustive Literature Search

Once criteria for inclusion of studies in a best-evidence synthesis have been established, it is incumbent upon the reviewer to locate every study ever conducted that meets these criteria. Books on meta-analysis (e.g., Cooper, 1984; Light & Pillemer, 1984) give useful suggestions for conducting literature searches using ERIC, Psychological Abstracts, Social Science Citation Index, and bibliographies of other reviews or meta-analyses, among other sources. In some cases, it is necessary to write to authors to request means and standard deviations or other information necessary to understand some aspect of a study. It is particularly important to locate all studies cited by previous reviewers to assure the reader that any differences in conclusions between reviewers are not simply due to differences in the pool of studies located.

Computation of Effect Sizes

In general, effect sizes should be computed as suggested by Glass et al. (1981), with a correction for sam-

ple size devised by Hedges (1981; Hedges & Olkin, 1985). The Hedges procedure produces an unbiased estimate of effect size, reducing estimates from studies with total *N*'s (experimental plus control) less than 50.

There are many statistical issues that are important in computing and understanding effect sizes, and many of these have important substantive implications. For example, there are questions of how to interpret gain scores or posttests adjusted for covariates, how to deal with unequal pretest scores in experimental and control groups, and how to deal with aggregated data (e.g., class or school means). Readers interested in statistical issues should refer to the excellent books on the conduct of quantitative syntheses (e.g., Cooper, 1984; Glass et al., 1981; Hedges & Olkin, 1985; Hunter et al., 1982; Rosenthal, 1984).

Averaging effect sizes within studies. Since many studies report a large number of effects, it may be important to compute averages of some effect sizes across particular subsets of comparisons. The amount of averaging to be done depends on the purpose and focus of the best-evidence synthesis. For example, in a general review of the effects of ability grouping on achievement, different measures of reading and language arts might be averaged. However, in a best-evidence synthesis of research on specific reading strategies, we would want to preserve information separately for reading comprehension, reading vocabulary, oral reading, language mechanics, and so on.

Similarly, in a review of effects of computer-assisted instruction we might average effects for students of different ethnicities, but in a review of compensatory education, separate effects for different ethnic groups might be preserved. However, when pooling effect sizes across studies, each study (or each experimental-control comparison) must count as one observation with effect sizes from similar measures averaged as appropriate. To count each dependent measure as a separate effect size for pooling purposes, as recommended by Glass et al. (1981), creates serious problems as

it gives too much weight to studies with large numbers of measures and comparisons and violates assumptions of independence of data points in any statistical analyses (see Bangert-Drowns, 1986).

Table of Study Characteristics and Effect Sizes

No matter how extensive the literature reviewed, all studies should be listed in a table specifying major design and setting variables and effect sizes for principal studies. This table should include the names of the studies, sample size, duration, research design, subject matter, grade levels, treatments compared, and effect size(s). Other information important in a particular area of research might also be included. For example, the table might indicate which effects were statistically significant in the original research. This table is essential not only in summarizing all pertinent information, but also in making it easier to check the review's procedures and conclusions against the original research on which it was based.

In the table of study characteristics and effects sizes, results from studies for which effect sizes could not be computed may be represented as "+" (statistically significant-positive), "0" (no significant differences), or "-" (statistically significant-negative).

For examples of tables of study characteristics and effect sizes, see Willig (1985), Schlaefli, Rest and Thoma (1985), Kulik and Kulik (1984), and Slavin (1986).

Pooling of Effect Sizes

When there are many studies high in internal and external validity on a well defined topic, pooling (averaging) effect sizes across the various studies may be done. For example, let's say we located a dozen studies of Treatment X in which experimental and control students (or classes) were randomly assigned to treatment groups, the treatment was applied for at least 3 weeks, and fair achievement tests equally responsive to the curriculum taught in the experimental and control groups were used. In this case, we might pool the effect sizes by computing a median across the 12 studies. Medians are preferable to

means because they are minimally influenced by anomalous outliers frequently seen in meta-analyses.

In pooling effect sizes, the reviewer must be careful "not to quantitatively combine studies at a broader conceptual level than the readers would find useful" (Cooper, 1984, p. 82). For example, in a quantitative synthesis by Lysakowski and Walberg (1982), it was not useful to pool across studies of cues, participation, and corrective feedback, as these topics together do not form a single well-defined category (see Slavin, 1984).

Pooled effect sizes should be reported as adjuncts to the literature review, not its primary outcome. Pooling and statistical comparisons must be guided by substantive, methodological, and theoretical considerations, not conducted wholesale and interpreted according to statistical criteria alone. For example, many meta-analyses routinely test for differences among effect sizes according to year of publication, a criterion that may be important in some literatures but is meaningless in others, while ignoring more theoretically or methodologically important comparisons (such as plausible interactions among study features).

Pooled effect sizes should never be treated as the final word on a subject. If pooled effects are markedly different from those of two or three especially well designed studies, this discrepancy should be explained. Pooling has value simply in describing the central tendency of several effects that clearly tend in the same direction. When effects are diverse, or the number of methodologically adequate, germane articles is small, pooling should not be done. Hedges and Olkin (1985) have described statistical procedures for testing sets of effect sizes for homogeneity, and these may be useful in determining whether or not pooling is indicated. However, decisions about which studies to include in a particular category should be based primarily on substantive, not statistical criteria.

Literature Review

The selection of studies, computation of effect sizes, and pooling de-

scribed above are only a preliminary to the main task of a best-evidence synthesis: the literature review itself. It is in the literature review section that best-evidence synthesis least resembles meta-analysis. For example, some quantitative syntheses do use a priori selection, do present tables of study characteristics and effect size, and do follow other procedures recommended for best-evidence synthesis, but it is very unusual for a quantitative synthesis to discuss more than two or three individual studies or to examine a literature with the care typical of the best narrative reviews.

There are no formal guidelines or mechanistic procedures for conducting a literature review in a best-evidence synthesis; it is up to the reviewer to make sense out of the best available evidence.

Formats for Best-Evidence Syntheses

No rigid formula for presenting best-evidence syntheses can be prescribed, as formats must be adapted to the literature being reviewed. However, one suggestion for a general format is presented below. Also, see Slavin (1986) for an example of a best-evidence synthesis.

Introduction. The introduction to a best-evidence synthesis will closely resemble introductions to traditional narrative reviews. The area being studied is introduced, key terms and concepts are defined, and the previous literature, particularly earlier reviews and meta-analyses, is discussed.

Methods. In a best-evidence synthesis, the methods section serves primarily to describe how studies were selected for inclusion in the review. The methods section might consist of the following three subsections.

Best-Evidence Criteria describes and justifies the study selection criteria employed. Clear, quantifiable criteria must be specified, not global ratings of methodological adequacy. Stringent criteria for germaneness should be applied (e.g., studies of individualized instruction in mathematics that took place over periods of at least 8 weeks in elementary schools, using mathematics achievement mea-

asures not specifically keyed to the material being studied in the experimental classes). Among germane studies, criteria for methodological adequacy are established, focusing on avoidance of systematic bias (e.g., use of random assignment or matching with evidence of initial equality), sample size (e.g., at least four classes in experimental and control groups), and external validity (e.g., treatment duration of at least eight weeks). The literature search procedure should be described in enough detail that the reader could theoretically regenerate an identical set of articles. A section titled *Studies Selected* might describe the set of studies that will constitute the synthesis, while a section on *Studies Not Selected* characterizes studies not included in the synthesis, in particular describing excluded studies that were included in others' reviews and studies that "just missed" being included.

Literature Synthesis. The real meat of the best-evidence synthesis is in the *Literature Synthesis* section. This is where the research evidence is actually reviewed. This section would first present and discuss the table of study characteristics and effect sizes and discuss any issues related to the table and its contents. If pooling is seen as appropriate, the results of the pooling are described; otherwise, the rationale for not pooling is presented.

In a meta-analysis, the presentation of the "results" is essentially the end point of the review. In a best-evidence synthesis, the table of study characteristics and effect sizes and the results of any pooling are simply a point of departure for an intelligent, critical examination of the literature (see Light & Pillemer, 1984). In the *Literature Synthesis* section, critical studies should be described and important conceptual and methodological issues should be explored. A best-evidence synthesis should not read like an annotated bibliography, but should use the evidence at hand to answer important questions about effects of various treatments, possible conditioning or mediating variables, and so on. When conclusions are suggested, they must be justified in light of the available evidence, but also the *contrary* evidence should be

discussed. Effect size information may be incorporated in the *Literature Synthesis*, as in the following example:

"Katz and Jammer (19XX) found significantly higher achievement in project classes than in control classes on mathematics computations (ES = .45) and concepts (ES = .31), but not on applications (ES = .02)."

In general, the "best-evidence" studies should be described with particular attention to studies with outstanding features, unusually high or low effect sizes, or important additional data. Studies that meet standards of germaneness and methodological adequacy but do not yield effect size data should be discussed on the same basis as those that do yield effect size data. Studies excluded from the main synthesis may be brought in to illustrate particular points or to provide additional evidence on a secondary issue. Except for the references to effect sizes, the bulk of the *Literature Synthesis* should look much like the main body of any narrative literature review.

One useful activity in many best-evidence syntheses is to compare review-generated and study-generated evidence (see Cooper, 1984). Review-generated evidence results from comparisons of outcomes in studies falling into different categories, while study-generated evidence relates to comparisons made within the same studies. For example, a reviewer might find an average effect size of 1.0 in methodologically adequate studies of Treatment X, and 0.5 in similar studies of Treatment Y and conclude that Treatment X is more effective than Treatment Y. However, this is not necessarily so, as other factors that are systematically different in studies of the two treatments could account for the apparent difference. This issue could be substantially informed by examination of studies that specifically compared treatments X and Y. If such studies exist and are of good quality, they would constitute the best evidence for the comparison of the treatments. Review-generated evidence can be useful in *suggesting* comparisons to be sought within studies, and may

often be the only available evidence on a topic, but is rarely conclusive in itself.

Conclusions. One purpose of any literature review is to summarize the findings from large literatures to give readers some indication of where the weight of the evidence lies. A best-evidence synthesis should produce and defend conclusions based on the best available evidence; or in some cases may conclude that the evidence currently available does not allow for any conclusions.

Summary

The advent of meta-analysis has had an important positive impact on research synthesis in reopening the question of how best to summarize the results of large literatures and providing statistical procedures for computation of effect size, a common metric of treatment effects. It is difficult to justify a return to reviews with arbitrary study selection procedures and reliance on statistical significance as the only criterion for treatment effects. Yet in actual practice (at least in education), meta-analysis has produced serious errors (see Slavin, 1984).

This paper proposes one means, best-evidence synthesis, of combining the strengths of meta-analytic and traditional reviews. Best-evidence synthesis incorporates the quantification and systematic literature search methods of meta-analysis with the detailed analysis of critical issues and study characteristics of the best traditional reviews in an attempt to provide a thorough and unbiased means of synthesizing research and providing clear and useful conclusions. No review procedure can make errors impossible or eliminate any chance that reviewers' biases will affect the conclusions drawn. It may be that applications of the procedures proposed in this paper will still lead to errors as serious as those often found in meta-analytic and traditional reviews. However, applications of best-evidence synthesis should at least make review procedures clear to the reader and should provide the reader with enough information about the primary research on which the review is based to reach independent conclusions.

References

- Bangert-Drowns, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, 99, 388-399.
- Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. *Journal of Special Education*, 14, 295-309.
- Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare.
- Cook, T., & Leviton, L. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, 48, 449-472.
- Cooper, H.M. (1984). *The integrative research review: A systematic approach*. Beverly Hills, CA: Sage.
- Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, 33, 517.
- Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research*, 5, 3-8.
- Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Glass, G., Cohen, L., Smith, M. L. & Filby, N. (1982). *School class size*. Beverly Hills, CA: Sage.
- Glass, G. & Stanley, J.C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.
- Goldstein, H., Moss, J., & Jordan, J. (1966). *The efficacy of special class training on the development of mentally retarded children* (Cooperative Research Project no. 619). Washington, DC: U.S. Office of Education.
- Gottfredson, S. (1978). Evaluating psychological research reports. *American Psychologist*, 33, 920-934.
- Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107-128.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Jackson, G. B. (1980). Methods for integrative reviews. *Review of Educational Research*, 50, 438-460.
- Jackson, S.E. (1984, August). *Can meta-analysis be used for theory development in organizational psychology?* Paper presented at the annual convention of the American Psychological Association, Toronto.
- Kulik, J. A., & Kulik, C. L. (1984). Effects of accelerated instruction on students. *Review of Educational Research*, 54, 409-425.
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Lysakowski, R., & Walberg, H. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal*, 19, 559-578.
- Madden, N. A., & Slavin, R. E. (1983). Mainstreaming students with mild academic handicaps: Academic and social outcomes. *Review of Educational Research*, 53, 519-569.
- Marsh, H., & Ball, S. (1981). Interjudgmental reliability of reviews for the *Journal of Educational Psychology*. *Journal of Educational Psychology*, 73, 872-880.
- Mintz, J. (1983). Integrating research evidence: A commentary on meta-analysis. *Journal of Consulting and Clinical Psychology*, 51, 71-75.
- Moody, W. B., Bausell, R. B., & Jenkins, J. R. (1973). The effect of class size on the learning of mathematics: A parametric study with fourth grade students. *Journal for Research in Mathematics Education*, 4, 170-176.
- Ottensbacher, R.J., & Cooper, H.M. (1983). Drug treatment of hyperactivity in children. *Developmental Medicine and Child Neurology*, 25, 358-366.
- Page, E. (1975). Statistically recapturing the richness within the classroom. *Psychology in the Schools*, 12, 339-344.
- Peters, D., & Ceci, S. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *The Behavioral and Brain Sciences*, 5, 187-255.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.
- Scarr, S., & Weber, B. (1978). The reliability of reviews for the *American Psychologist*. *American Psychologist*, 33, 935.
- Schlaefli, A., Rest, J. R., & Thoma, S. J. (1985). Does moral education improve moral judgment? A meta-analysis of intervention studies using the defining issues test. *Review of Educational Research*, 55, 319-352.
- Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher*, 13 (8), 6-15, 24-27.
- Slavin R. E. (1985, March). *Quantitative review*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Slavin, R. E. (1986). *Ability grouping and student achievement in elementary schools: A best-evidence synthesis* (Tech. Rep. No. 1). Baltimore, MD: Center for Research on Elementary and Middle Schools, Johns Hopkins University.
- Verducci, F. (1969). Effects of class size on the learning of a motor skill. *Research Quarterly*, 40, 391-395.
- Waxman, H., & Walberg, H. (1982). The relation of teaching and learning: A review of reviews of process-product research. *Contemporary Education Review*, 1, 103-120.
- Waxman, H.C., Wang, M. C., Anderson, K. A., & Walberg, H.J. (1985). Adaptive education and student outcomes: A quantitative synthesis. *Journal of Educational Research*, 78, 228-236.
- Willig, A.C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research*, 55, 269-317.
- Wilson, G.T., & Rachman, S.J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology*, 51, 54-64.

Clarification

On p. 21 of the October *ER*, an alternate sigma (σ) symbol was used in the equations below. For those readers who may have been confused by this symbol, the equations now appear with the usual sigma (σ) symbol. In addition, an extraneous subscript p appeared in the first equation.

$$E(e_i) = 0, \sigma_{e_i}^2 = (1 - \rho_i^2)^2 / (N - 1), \sigma_r^2 = \sigma_{\rho_i}^2 + \sigma_{\rho_e}^2$$

$$" \sigma_{\rho_i}^2 = \sigma_{\rho_i}^2 / r_{xx} r_{yy} " \text{ (p. 56); } " E(d) = \delta " \text{ (p. 101);}$$

$$" \sigma_{\rho_e}^2 = 4(1 + \delta^2/8)/N " \text{ (p. 101).}$$