DEVELOP PERFORMANCE METRICS THAT DON'T RELY ON SURVEYS

George Beam

Associate Professor
Department of Public Administration
University of Illinois at Chicago
Chicago, Il 60607-7064

Summary We need to develop performance metrics that don't rely on surveys because answers to questions are unreliable. When you only have answers it's impossible to know which, if any, are correct or incorrect. Answers are made unreliable by respondents and by asking. Reliable performance metrics are acquired by using "proper"—as I name them—methods of data collection (observation and document analysis) and research designs (experiments, multiple sources, formal models, and comparison). These methods of data collection and these research designs—when used to generate information about behaviors of personnel and results of individuals, groups, and organizational subunits—produce reliable performance metrics.

Keywords   metrics, performance, surveys, observation, experiments, documents, models

Text

        We need to develop performance metrics that don't rely on surveys (defined here as

procedures or instruments that ask questions of respondents) because answers to questions are not

reliable. This is The Problem with surveys—with asking—namely, when you only have answers

to questions it's impossible to know if one or more of them are correct or accurate (Category One

Answers) or are incorrect or inaccurate (Category Two Answers). The unreliability of answers to

questions is not just my view or opinion; actually, The Problem with asking has been "well

documented" (Ketokivi and Schroeder, 2004, p. 251) and widely acknowledged.

        The *only* way to know if an answer is correct or accurate is to check, or verify, it with

information from two or more additional non-asking sources of information; say, from

observation, experimentation, and documents. Those who rely on surveys, interviews, or any other variety of the asking method—I call them, "askers"—do not have information from non-asking sources and, therefore, they have The Problem; they have unreliable information.

Askers do not accept these definitions of "unreliable" and "reliable". They contend that asking produces reliable information to the extent it produces consistent, reproducible, results or scores. But this notion of reliability may "`just represent everybody together getting the same wrong answer'" (Howwitz and Wakefield, 2007, as quoted in Crews, 2007, Dec. 6, p. 14). Or the same *right* answer. That's The Problem with asking; when all you have are answers to questions, you can't tell whether or not one or more of them are right or wrong.

Answers to questions are made unreliable by respondents and by asking. Because there are no answers without respondents and asking—and because both respondents and asking make answers unreliable—answers are, in this sense, inherently unreliable.

Respondents Make Answers Unreliable

Respondents make answers unreliable, skewing every answer they give into Category One or Category Two, because (1) they sometimes lie, (2) they may not have relevant and correct information, (3) their values and norms affect answers, and because (4) their interest in, and (5) sensitivity to, the questions put to them bias their answers. Moreover, (6) there are many other ways respondents contribute to The Problem; some mentioned in the following pages.

Virtually everyone knows that people, when asked, sometimes and, depending upon circumstances—for example, who's asking whom, about what—often lie. Lying is a daily affair that permeates all of life. And no topic—including performance—is immune to prevarication (Huber and Power, 1985; Ketokive and Schroeder, 2004).

Another reason why organizational personnel and other respondents make answers

unreliable is that some of them do not have relevant and accurate information (Huber and Power, 1985). For example, the events, decisions, and developments related to performance "frequently involve multiple participants, a significant number of whom may lack full information" (Huber and Power, 1985, p. 173). Respondents to surveys (and everyone else) are, to various degrees, misinformed and uninformed about many matters. Rather than admit ignorance, some guess and, more than a few, incorrectly. Others devoid of appropriate information, but thinking otherwise, also give answers that are off the mark.

In addition, respondents contribute to The Problem because they skew their replies to correspond to social values and norms, as well as to the priorities and perspectives associated with their organizational positions (Huber and Power, 1985; Ketoivi and Schroeder, 2004). When asked about their behavior, responders affirm they acted consistent with social and organizational norms when they haven't, and deny performing, or participating in, undesirable behaviors when they have (Parry and Crossley, 1950; Phillips and Clancy, 1970; Riesman and Benny, 1956; Silver, Ambramson, and Anderson, 1986; Sproull and Kiesler, 1991).

The interest respondents have in question topics makes their answers unreliable. People who appear to be interested in a topic respond at a higher rate and give different answers than those who, presumably, have less, or no, interest (Dolsen and Machlis, 1991; Jurkiewicz and Nichols, 2002). Thus, contents of questions—that is, the issues or topics investigated—can generate "selection bias", skewing results according to the concerns or interests of responders.

The extent to which topics are considered sensitive by respondents makes answers to questions unreliable. The greater the sensitivity of question topics, the greater the effects on rates of response and on reports of the behaviors and attitudes investigated. Moreover, sensitive question topics contribute to The Problem because what's considered sensitive varies by

respondents' childhood experiences, peer and professional socialization, present and past socioeconomic statuses, organizational position and functions, plans for the future, and so on. Consequently, *any* question topic can be sensitive, and this means any topic or issue can skew answers into Category One or Category Two.

There are many other ways respondents contribute to The Problem. For example, those who are overly committed to advancing their own convictions disregard the questions put to them in favor of making their own points. Other responders contribute to The Problem by having another person answer for them. Claiming to be "too busy", organizational personnel in higher positions pass questionnaires to those at lower levels. Some husbands oblige their wives to handle these inconveniences. Also, it's not uncommon for respondents to contribute to The Problem by marking middle or average positions on Likert scales, rather than carefully considering which position on the scale corresponds to their actual opinion or behavior.

Asking Makes Answers Unreliable

Asking makes answers unreliable because answers are skewed by (1) specific characteristics of asking instruments, including wording of questions, (2) peculiarities of settings or environments in which questions are asked and answers given, and by the (3) attributes and behaviors of askers. "[A]n unbiased survey instrument"(Goldstein, 2010, p. 1) is an impossibility.

The wording of questions in asking instruments is the use of specific words and groups of words (phrases). Wording identifies the topic under investigation and presents questions in a particular style, diction, and idiom. It's important to realize that it's not possible to word a survey or any other asking instrument so that questions—i.e., the words that constitute the questions— have no effect on answers. As Schuman (2008) points out, "[a]nswers depend on questions" (p.

13).

Moreover, question wording biases answers because much of every language is ambiguous. Documented instances of ambiguous words in questions include: "needed services" (Asher, 2004, p. 168), violence (Gorner and Dell'Angela, Aug. 1, 2001), quality (Executives believe quality contributes to bottom line, April 2004), exercise (Asher, 2004), "`welfare'", "`big business'", "`civil rights'", "`profits'", "`energy crisis'", "`big government'", (Sudman, Bradburn, and Wansink, 2004, p. 120), and eldercare (*Exploring employee utilization of employer-sponsored eldercare programs*, 2003). Also, small changes in questions wording—such as "substituting `helping the poor' for `welfare'" (Vasu, Stewart, and Garson, 1998, p. 164)—affect answers, making them unreliable. And it's unlikely that the words "strongly agree" or "above industry average" are interpreted by all respondents "in a similar, consistent and comparable manner" (Ketokivi and Schroeder, 2004, p. 251). To the extent answers are affected by the wording of questions, rather than by the phenomena investigated—such as the effects of a bundle of human resource programs on performance, actual opinions about the values of a candidate for promotion, and so on—wording in asking instruments erodes the reliability of answers and, thereby, contributes to The Problem.

Another characteristic of asking instruments that makes them produce unreliable information is that the questions that constitute instruments are symbolic and unrealistic, and this means asking instruments tend to produce symbolic and unrealistic answers. Words are symbols; verbal symbols that do not necessarily capture—sufficiently, or at all—the realities for which they are supposed to be referents. The word, "manager", for instance, *represents* a particular type of human being, but the word, "manager", is not a particular, real person in a particular, actual place with the unique characteristics of that particular manager in that particular place or situation.

Notoriously symbolic and unrealistic are asking instruments querying people about their values. Such questions do not take into account the fact that people hold numerous values and, in many real-life situations, some values conflict with others. Moreover, the extent of commitment to any one value and its effect on what a person says and does depends on how a particular value relates to other priorities in specific real-life situations.

Another indication that asking instruments produce unreliable answers is that when identical questions are put to the same populations or samples by different instruments, each instrument produces different response rates and/or different substantive answers than any other instrument (Asher, *Polling and the Public*, 2004; Hochstim, 1967; Hyman, 1944-45; Katz, 1942; Linn, 1965; Tourangeau, Jobe, Pratt, and Rasinski, 1997).

Also, asking instruments contribute to The Problem because they often generate conflicting or inconsistent answers from individual respondents. Instruments that contain numerous questions on a particular topic—such as, illicit drug use, voting, abortions, or quality of life usually produce answers that affirm the behavior or opinion investigated, whereas other responses from the same respondents imply, or explicitly state, the opposite (Bennett and DiLorenzo, 1992; Miller, 1997).

Reinterviewing instruments—such as panel studies—often generate inconsistent answers. Respondents give answers in repeat interviews that conflict with their responses to the same questions in initial asking sessions (Brehm, 1994; Fendrich and Vaughn, 1994).

Another characteristic of asking instruments that contributes to The Problem is that they produce nonresponse and, usually, a lot of it. Nonresponse can make answers unreliable because nonrespondents would likely have given different answers than did respondents. Thus, the results of survey research efforts would be different if nonrespondents had responded.   (This is

especially true for telephone surveys because responders to telephone surveys are different people than responders.) As a consequence of nonresponse, important characteristics of samples often are significantly overestimated and/or underestimated.

Asking instruments also tend to produce unrepresentative results and, in that respect, further contribute to The Problem. Results of asking instruments are unrepresentative when answers are from respondents who are not representative of the whole population that is investigated; that is, from respondents who do not share the same demographic or other relevant characteristics of the whole population. Those with the same demographic characteristics, the same life experiences, the same organizational position and functions, and so on, tend to say the same or similar words in response to a particular question or set of questions; whereas those experiencing different cultures, backgrounds, situations, positions or offices, and events answer accordingly and differently. Thus, responses of a representative sample can be generalized to the whole population, but answers of an unrepresentative sample cannot. The answers of an unrepresentative sample merely indicate what that particular group said and, in that sense, they are unreliable indicators of what the whole group might have said.

Formats of asking instruments affect results—both response rates and content of answers—to various degrees; thus, instrument formats contribute to The Problem (Couper, 2000). Formats are: (1) physical features of instruments, (2) structures of questions (for example, open-ended, and fixed response questions), and (3) patterns in which questions are related to each other (for instance, placing questions about personal matters; such as gender and income, before or after questions about the topic or issue being investigated).

It's also the case that asking instruments produce unreliable results, skewing answers into Category One or Category Two, because the effects of various components of instruments on

answers—such as question wording and instrument format—are tangled, making it impossible to identify the specific effects that a particular instrument component has on answers (Tourangeau, et al., 1997). Moreover, the effects of instrument components on answers also are mixed with the effects of asking settings and askers themselves. To the extent effects of components of instruments, settings, and askers are tangled, it cannot be known what component of the survey, interview, or the like is causing what amount of response bias and, therefore askers cannot adjust for it. This is another reason survey researchers and other askers do not have, and cannot acquire, reliable information.

Answers to questions are, in addition, made unreliable because answers are affected by the settings in which asking occurs. Generally, each asking setting or situation generates results (response rates and contents of answers) that are different than when the same or similar questions are asked in different settings.

There are two basic types of settings or situations that affect responses: (1) societal settings and, (2) immediate settings. Societal settings are the cultures experienced by respondents, and that includes social values, perspectives, illusions, religions, and understandings of organizations (processes, structures, and personnel), politics, and economics. Also included in respondents' societal settings are their positions in society—such as socioeconomic position, marital position or status, and so on—and the norms and beliefs associated with these positions. Via socialization by parents, schools, the mass media, and colleagues, respondents take as their own the cultural and positional preferences, priorities, and outlooks they experience, and form their answers accordingly.

Immediate settings are the specific places where asking and answering occur; such as respondents' workplaces and homes, offices and rooms at places of business where consultants

interview personnel, and rooms in survey research centers in which focus groups are administered.

Immediate asking settings are "contaminat[ed]", in that components of settings skew responses (Anderson and Silver, 1987, p. 539). These contaminations are "forces that affect" what people say about their attitudes and behaviors (Schuman and Johnson, 1976, p. 191). Many components of immediate asking settings—such as the presence of third parties, as well as the design, appearance, and comfort-level of buildings and rooms in which questions are asked—bias answers obtained in those settings (O'Rourke, 2000; Zanes and Matsoukas, 1979). A voice recorder during a face-to-face interview contaminates the setting because its existence makes respondents more cautious or careful in answering than would be the case if a recorder were not used.

Immediate asking settings also contribute to The Problem because different settings generate different answers to the same questions. In addition, many immediate asking settings skew answers because they are unrealistic, or "unnatural" (Morgan, 1997, p. 8); that is, the environments in which questions are posed are different in many significant respects than the situations or settings in which investigated opinions and actions are actually formed, stated, and performed. Specifically, and significantly, immediate asking situations are absent the social and organizational forces that shape respondents' words and actions in everyday life (Fendrich, 1967). Because asking situations do not and, usually, cannot, replicate situations in which investigated opinions, intentions, beliefs, knowledge, and actions have been, will become, or are, operative, answers obtained in asking settings often are different than answers in actual, lived situations and, moreover, do not indicate what respondents would say or do in real life circumstances.

Another characteristic of immediate asking settings that affects answers, and thereby

contributes to The Problem, is that asking settings are separated in time from the phenomena being investigated. Thus, those asking for information are dependent upon respondents' memory (Ericsson and Simon, 1993). Askers, therefore, must assume that respondents: (1) have actually experienced the phenomenon being investigated, (2) have retained the experience in their memory, and (3) have recalled the experience. There is room for error in all three memory-elements of answers.

Additionally, both societal and immediate settings contribute to The Problem because each generates emotions that affect answers to questions (Turner and Krauss, 1978). Moreover, the effects of societal and immediate asking settings on response rates and contents of answers prevent (as do the unrepresentative results discussed above) the generalization of asking results to other situations or settings; that is, answers obtained in one setting do not indicate what answers will be to the same questions in any other setting.

Askers (as well as instruments and settings) contribute to The Problem because askers are, as I call them, "stimulators" and, as such, they "cue" (Cannell and Kahn, 1968, p. 550; Fendrich and Vaughn, 1994, p. 119) and induce response rates and reports by their specific styles of behavior while asking; by their individual personal attributes, such as their judgments when coding responses; and by their particular experiences, competencies, ethnicity, socioeconomic features, gender, and age (Ferber and Wales, 1952; Cahalan, Tamulonis, and Verner, 1947; Rice, 1929).

Also, askers skew answers because they do "everything possible" (Hochstim, 1967, p. 977) to obtain higher response rates and more reports. In everyone of these efforts—and this includes: wording and phrasing questions, defining and re-defining words, controlling the development and administration of interviews and other asking efforts, persisting (callbacks, etc.),

shaming, intimidating, bribing (incentives), recruiting respondents, promising anonymity and confidentiality, probing, prompting, and using props—askers affect results and, thereby, contribute to The Problem.

Asking also contributes to The Problem because the effects of asker behaviors and attributes are tangled. The effects of askers' body language on answers are comingled with the effects of askers' age and/or gender. The effects of interviewers' education and age are tangled with the effects of interviewers' training (Andersen and Olsen, 2002).

In addition, asking makes answers unreliable because the effects of asker characteristics (e.g., gender, age, and ethnicity) on answers are tangled with effects produced by other components of asking efforts; such as asking instruments and settings. Moreover, the influence of the behavior of interviewers is "confounded" with the influence of bribes/incentives (Willimack, Schuman, Pennell, and Lepkowski, 1995, p. 81). These entanglements and mixings make it impossible to identify which component of asking has what effect on results. Thus, whether or not one or more answers are in Category One or Category Two cannot be established.

The effects of instruments, settings, and askers on answers mean that answers to questions about performance—or anything else—are not "out there" to be found, nor are they generated from within respondents, nor are answers voluntary. Answers are responses to the stimuli and reinforcements of asking. Answers are produced or "manufactured" (Prior, 2003, p. 44) by components of the asking method; that is, by asking instruments, asking settings, and askers themselves. No matter how many identical, similar, and/or different questions are asked, regardless of which asking instruments are used, despite the settings in which questions are asked and answers given, irrespective of the attributes and behaviors of askers, answers to questions remain skewed or biased by the components of the asking method and, thus, unreliable.

Here's my summary of The Problem with asking: Although every answer—whether about the objective or subjective aspects of performance or any other phenomenon—is skewed by both respondents and by asking, it's possible that answers are skewed so that they are correct or true (Category One Answers), *and* it's possible answers are skewed so that they are incorrect or false (Category Two Answers). This is to say, *all* answers to questions *are* skewed, *and* answers are either correct or incorrect. When you only have answers to questions, you have The Problem because it's not possible to know if the answers are correct or incorrect.

Proper Methods of Data Collection and Proper Research Designs

Rather than asking to find out about performance—or anything else—we should use what I call "proper" methods of data collection (observation and document analysis) and proper research designs (experimentation, multiple sources, formal models, and comparison). With the use of proper methods and research designs, we can obtain accurate data that describes existing performance, as well as information that will help us increase individual, subunit, and organizational output.

Answers to questions about performance are not needed because information or data about performance can be acquired by observation. Specifically, we should observe "`something doing'"; that is, "observ[e]. . . . actually performed . . . activities" (Bentley, 2008, pp. 176, 187, 180; Mintzberg, 1973, pp. 226-29). Our understanding of organizations—and what goes on within them, such as decision making and performance—"should be connected to empirical observations of what happens in [organizations]" (Cyert and March, 1992, p. xi). Consulting firm, Aubrey Daniels International, Inc., bases its recommendations for increasing performance on observations (Daniels and Daniels, 2004).

We can also obtain reliable performance data by observing the "physical traces surviving from past behavior" (Webb, Campbell, Schwartz, and Sechrest, 2000, p. 35).   For example, every interaction with a Personal Computer (PC) leaves trace data behind. Thus, PCs can be investigated for traces of activity that demonstrate performance.

It's also the case that reliable data about performance—and, in some instances, indications for improving it—are produced by experiments (Bryman, 1989; Ganster, 1980; Orpen, 1979). In an experiment, one or more changes are made to an organization's processes, structure, or personnel in order to identify and explain the changes that may occur (Montgomery, 2001). Typically, in a performance measurement experiment, the investigator administers to one or more groups of people, organizational subunits, or assembly lines, one or more interventions or treatments (such as various work group structures, or types of equipment) to measure what effect, if any, the treatment(s) has on individual and/or unit productivity. In other words, experiments attempt to establish causality: in this example, does the intervention or treatment (independent variable) cause a change of direction in productivity (dependent variable)?

The use of multiple sources is another research design for generating reliable performance data. Actually, multiple sources of data about performance—and all other objects of investigation—should be considered essential in every investigation because the results of any effort to obtain reliable information—including data acquired by proper methods and research designs—might be biased or incomplete. It's possible that even the most highly trained observers under the best conditions will miss at least a few relevant phenomena. Or, certain activities and/or indicators of performance will be incompletely perceived and, thus, not properly accounted for in experiments. Moreover, and invariably, all measurements are less than 100 percent accurate.

The well-recognized procedure, or research design, that counters—but can never eliminate—limitations in results from any single method of data collection or research design, is generation of data, or measurements, from multiple sources. This can be done by the use two or more methods of data collection (such as, observation and document analysis), the use of two or more research designs (such as experiments and comparison), or the use of two or more combinations of data collection methods and research designs (Campbell and Fiske, 1959; Gorard, with Taylor, 2004; Sharpe and Koperwas, 2003; Webb, et al., 2000).

Constructing and testing formal models is another proper research design for generating performance data. Many models are constructed with computer software. A computer model is often referred to as a simulation, a simulation model, or a computer simulation (Bryman, 1989; Gilbert and Troitzsche, 1999).

A formal model is a set of simplified assumptions that describes what's being investigated—for example, the performance of an organization's payroll department—from which testable conclusions are deduced. Empirical support for the conclusions is support for the assumptions in the model. As a formal model's deduced conclusions and assumptions are empirically substantiated, "the model as a whole" (Cyert and March, 1992, p. 87) becomes a more complete explanation—in this example—of the payroll department and its output.

Document analysis is another proper method for collecting reliable information about performance (Bryman, 1989; Prior, 2003; Ventresca and Mohr, 2005). In one instance, data from a manufacturing plant's records, as well as information from work teams' meeting logs and training documents were used to assess the production of a manufacturing plant's work team (Banker, Field, Schroeder, and Sinha, 1996). Hendricks and Singhal (1997) "use publicly available accounting data to test for changes in operating performances that result from

implementing effective TQM programs" (p. 251). (Whenever possible, primary, rather than secondary, documents should be used because when initial or primary documents are restated, abbreviated or, in other ways, interpreted and presented, information and data often are lost, skewed, and/or misstated.)

Comparison is another research design that can produce reliable descriptions and explanations of performance. "Basic to scientific evidence . . .is the process of comparison, of recording differences, or of contrast. Any appearance of . . . intrinsic knowledge about singular isolated objects . . . is found to be illusionary upon analysis. Securing scientific evidence involves making at least one comparison" (Campbell and Stanley, 1963, p. 6).

In one type of comparative research design, investigators present what sociologist, Max Weber (1864-1920), termed an "ideal type" of the phenomenon being investigated; for example, bureaucracy. Then, the ideal type is compared or contrasted with actual instances of the phenomenon and, on the basis of the similarities and differences, researchers construct and test hypotheses that postulate explanations—causes—for the similarities and differences.

Another variety of comparative research design begins with the identification of similarities and differences between two or more actual instances of the phenomenon being investigated. Researchers interested in measuring and increasing production can compare, for example, the specific behaviors of personnel in two work groups, the structures and processes under which they operate, and the outputs of the two groups. Causal explanations for the differences and similarities are, then, hypothesized. Empirical evidence is acquired, and the hypothesized causal relationships are, or are not, supported.

The comparative research design is especially helpful when problems or issues appear to be unique or unprecedented. Most likely, others have dealt with similar issues, and comparison

often can lead to the identification of remedies that were not previously considered. Also, comparison can provide insights into relationships between organizational changes—for example, changes in formal structures and/or procedures—and variations in outcomes. Comparison of a number of case studies of a program that failed to increase performance can lead to one or more explanations for the failures. In one comparative study, medical school-based physicians were found to be less productive (in terms of patient revenue, and other outputs) than community-based physicians and, in this same study, reasons for the differences were identified (Serrin, 1999).

Ingenuity of the Shared Enterprise

To optimize the advance of knowledge about productivity—its causes, how to improve it, and related matters—requires more than the use of proper methods of data collection and proper research designs by individual investigators. It's also necessary that the whole community of researchers and practitioners work (1) to improve proper methods and research designs and (2) to develop additional procedures. This is to say, "individual creativity [will not] suffice. Threats to validity [in the measurement of performance] . . . never end. They are inevitable in the continuing search for knowledge. The effort to deal with them requires ingenuity, not just of the individual scholar but of the shared enterprise" (Webb, et al., 2000, p. xvi).

References

Anderson, B. A., and Silver, B. D. (1987). The validity of survey responses: Insights from interviews of married couples in a survey of Soviet emigrants. *Social Forces, 66*, 537-554.

Asher, H. B. (2004). *Polling and the public: What every citizen should know* (6th ed.) Washington, D.C.: CQ Press.

Banker, R. D., Field, J. M., Schroeder, R. G., and Sinha, K. K. (1966). Impact of work teams on manufacturing performance: A longitudinal field study. *Academy of Management Journal, 39*, 867-90.

Bennett, J. T., and DiLorenzo, T. J. (1992). *Official lies: How Washington misleads us*. Alexandria, VA: Groom.

Bentley, A. F. (2008). *The process of government: A study of social pressures*. New Brunswick: Transaction; first published in 1908.

Bradburn, N. M., Sudmann, S., and Wansink, B. (2004). *Asking questions*: *The definitive guide to questionnaire design: For market research, political polls, and social and health questionnaires* (Rev. ed.). San Francisco: Jossey-Bass.

Brehm, J. (1994). Stubbing our toes for a foot in the door? Prior contact, incentives and survey response". *International Journal of Public Opinion Research, 6*, 45-63.

Bryman, A. (1989). *Research Methods and Organizational Studies*. New York: Routledge.

Cahalan, D., Tamulonis, V., and Verner, H. W. (1947). Interviewer bias involved in certain types of opinion survey questions. *International Journal of Opinion and Attitude Research, 1*, 63-77.

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix". *Psychology Bulletin, 56*, 81-105.

Cannell, C. F., and Kahn, R. (1968). Interviewing. In G. Lindzey and E. Aronson (Eds.), *Handbook of social psychology*, 2nd ed. (pp.526-595). Reading, MA: Addison-Wesley.

Cook, S. W., and Selltiz, C. (1964). A multiple-indicator approach to attitude measurement. *Psychological Bulletin, 62*, 36-55.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*, 464-494.

Crews, F. (2007, Dec. 6). Talking back to prozac. *New York Review of Books, 54*(19), 10-14.

Cyert, R. M., and March, J. G. (1992). *A behavioral theory of the firm* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Daniels, A. C. and Daniels, J. E. (2004). *Performance management: Changing behavior that drives organizational effectiveness* (4th ed., Rev.). Atlanta, Georgia: Aubrey Daniels International.

Dolsen, D. E., and Machlis, G. E. (1991). Response rates and mail recreation survey results: How much is enough? *Journal of Leisure Research,23*, 272-277.

Ericsson, K. A., and Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT.

Executives believe quality contributes to bottom line. (April 2004). *PA Times,27*(4), 5.

Exploring employee utilization of employer-sponsored eldercare programs (May 2003). New York: New York Business Group on Health. Retrieved from http://www.nybgh.org/articles/eldercare.html

Fendrich, J. M. (1967). A study of the association among verbal attitudes, commitment and overt behavior in different experimental situations". *Social Forces, 45*, 347-355.

Fendrich, M., and Vaughn, C. M. (1994). Diminishing lifetime substance use over time: An inquiry into differential underreporting. *Public Opinion Quarterly, 58*, 96-123.

Ferber, R., and Wales, H. G. (1952). Detection and correction of interviewer bias. *Public Opinion Quarterly, 16*, 107-127.

Ganster, D. C. (1980), Individual differences and task design: laboratory experiment. *Organizational Behavior and Human Performance,* 26, 131-148.

Gilbert, N., and Troitzsche, K. G. (1999). *Simulations for the social scientist*. Buckingham: Open University.

Goldstein, S. D. (2010). *Superior customer satisfaction and loyalty: Engaging customers to drive performance*. Milwaukee, Wisconsin: ASQ Quality.

Gorard, S., with Taylor, C. (2004). *Combining methods in educational research*. Maidenhead, U.K.: Open University.

Gorner, P., and Dell'Angela, T. (Aug. 1, 2001). 1 in 5 girls say date got violent. *Chicago Tribune*, sec. 1, pp. 1, 16.

Hendricks, K. B., and Singhal, V. R. (1997). Does implementing an effective TQM program actually improve operating performance? Empirical evidence from firms that have won quality awards. *Management Science*, *43*, 1258-1274.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62*, 976-989.

Howwitz, A. V., and Wakefield, J. C. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. New York: Oxford University.

Huber, G. P., and Power, D. J. (1985). Retrospective reports of strategic-level managers: Guidelines for increasing their accuracy. *Strategic Management Journal,6*, 171-180.

Hyman, H. H. (1944-45). Do they tell the truth?" *Public Opinion Quarterly, 8I,* 557-559.

Jurkiewicz, C. L., and Nichols, K. L. (2002). Ethics education in the MPA curriculum: What difference does it make?" *Journal of Public Affairs Education, 8*, 103-114.

Katz, D. (1942) Do interviewers bias poll results? *Public Opinion Quarterly, 6*, 248-268.

Ketokivi, M. A., and Schroeder, R. G. (2004). Perceptual measures of performance: Fact or fiction? *Journal of Operations Management, 22*, 247-64.

Linn, L. S. (1965). Verbal attitudes and overt behavior: A study of racial discrimination. *Social Forces, 43*, 353-364.

Miller, P. V. (1997). Review: Is `up' right? The national household survey on drug abuse. *Public Opinion Quarterly, 61*, 627-641.

Mintzberg, H. (1973). *The nature of managerial work*. New York: Harper & Row.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). New Jersey: Wiley.

Morgan, D. L. (1997). *Focus groups as qualitative research* (2nd ed.) Thousand Oaks: Sage.

O'Rourke, D. (2000). An inquiry into declining RDD response rates: Part IV: lack of standardization. *Survey Research* Newsletter*, 31*(1), 1-2.

Orpen, C. (1979). The effects of job enrichment on employee satisfaction, motivation, involvement and performance: A field experiment. *Human Relations, 32*, 189-217.

Parry, H. J., and Crossley, H. M. (1950). Validity of responses to survey questions. *Public Opinion Quarterly, 14*, 61-80.

Phillips, D. L., and Clancy, K. J. (1970). Response biases in field studies of mental illness. *American Sociological Review, 35*, 503-515.

Prior, L. (2003). *Using documents in social research*. London: Sage.

Rice, S. A. (1929). Contagious bias in the interview: A methodological note. *American Journal of Sociology, 35*, 420-423.

Riesman, D., and Benny, M. (1956). Asking and answering. *Journal of Business, 29,* 225-236.

Schuman, H. (2008). *Method and meaning in polls and surveys*. Cambridge: Harvard University.

Schuman, H., and Johnson, M. P. (1976). Attitudes and behavior. *Annual Review of Sociology, 2*, 161-207.

Serrin, K.G. (1999). The cost of academic medicine: A physician productivity comparison of medical school-based and community-based orthopedic surgery groups. Washington, D.C.:

Association for Health Services Research. Meeting. Retrieved from
http://gateway.nlm.nih.gov/MeetingAbstracts/ma?f=102184944.html

Sharpe, T., and Koperwas, J. (2003). *Behavior and sequential analysis: Principles and
practice*. Thousand Oaks, CA: Sage.

Silver, B. D., Ambramson, P. R., and Anderson, B. A. (1986). The presence of others and
overreporting of voting in American national elections. *Public Opinion Quarterly, 50*,
228-239.

Sproull, L., and Kiesler, S. (1991). *Connections*: *New ways of working in the network
organization*. Cambridge, MA: MIT.

Tourangeau, R., Jobe, J.B., Pratt, W.F., and Rasinski, K. (1997). Design and results of the
women's health study". In L. Harrison and A.Hughes, (Eds.), *The validity of self-reported
drug use: Improving the accuracy of survey estimates* (pp. 344-365). Rockville, MD:
National Institute on Drug Abuse.

Turner, C. F., and Krauss, E. (1978). Fallible indicators of the subjective state of the nation.
American Psychologist, 33, 456-470.

Vasu, M. L., Stewart, D. W., and Garson, G. D. (1998). *Organizational behavior and public
management* (3rd ed.). New York: Marcel Dekker.

Ventresca, M. J., and Mohr, J. (2005). Archival research methods. In J. A. C. Baum, (Ed.),
Companion to organizations (pp. 805-828). Malden, MA.: Blackwell..

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (2000). *Unobtrusive measures*
(Rev. ed.). Thousand Oaks: Sage.

Willimack, D. K., Schuman, H., Pennell, B-H., and Lepkowski, J. M. (1995). Effects of a prepaid
nonmonetary incentive on response rates and response quality in a face-to-face survey.
Public Opinion Quarterly, 59, 78-92.

Zanes, A., and Matsoukas, E. (1979). Different settings, different results? A comparison of school
and home responses. *Public Opinion Quarterly, 43*, 550-557.

Copyright Information