POLICY: SOCIAL SCIENCE

Technology, Experimentation, and the Quality of Survey Data

David E. Bloom

Computers have wrought an extraordinary transformation in empirical social science research, having expanded enormously our capacity to collect, process, and analyze data. The consequent availability of unprecedentedly large amounts of data and the empirical results derived from them have made a huge contribution to our understanding of social conditions and human behavior. To the extent that informed decisions tend to be better decisions, computerization has given us a nearly ubiquitous tool to help improve the human condition.

Less widely recognized is computers' potential to improve the quality of the data we collect. For example, researchers now routinely have computers perform consistency checks by flagging out-of-bounds and internally inconsistent answers. Instead of resorting to the laborious and questionable practice of data cleaning, researchers can use computers to perform relevant checks on the spot and to probe for clarifications that result in clearer, more accurate data. In addition, via automating skip patterns, computers can ask more questions in a given amount of time and be more certain to follow the researchers' intended logic.

Interviewing by computer

The accompanying piece by Turner *et al.* (*1*) provides an ingenious demonstration of a new way in which computers can improve data quality. Using a cleverly designed experiment, they show that in relation to sensitive issues involving adolescent sexual behavior, drug use, and violence, computerized data collection may yield more accurate data than those gathered by self-administered, nonelectronic questionnaires. Because this finding calls into question the validity of much information that has been collected and used for public and private decisions in the past, it is profoundly important.

Turner *et al.* use a combination of laptop computers and audio-CASI (audio-enhanced, computer-assisted self-interviewing) technology to overcome two fundamental problems in the collection of high-quality survey data. The first problem concerns respondents' incentives to misstate the answers to questions that deal with issues they regard as sensitive. These incentives arise when survey respondents feel uncomfortable about answering particular questions aloud in the survey setting, which is typically their home, in the presence of others, including the interviewer. Researchers often address this problem by giving respondents the subset of sensitive questions in written form and asking them to record their answers in silence, thereby preserving, although sometimes only partially, the sense of privacy that presumably promotes honest answers. The second problem, which pertains only to the use of self-administered paper questionnaires, arises because of some respondents' limited literacy and their lingering concerns that their answers are not totally private.

The use of audio-CASI diminishes the significant barriers that these problems impose on the reliability of survey results. Turner *et al.*'s work provides persuasive evidence that the combined problems of sensitivity and literacy matter considerably, and that audio-CASI provides a possible way to overcome them. Audio-CASI may be particularly well suited to collecting data in inner-city settings and in developing countries, where overcrowded living conditions typically prevail, where literacy is relatively low, and where some of the behaviors in question may be particularly pronounced.

Audio-CASI removes all differences in how questions are delivered to respondents. It eliminates the effects of variations in intonation and speed of delivery, body language, and other subliminal signals. Presumably, it also reduces the well-documented effects of an interviewer's race, gender, and age on survey responses. Similarly, by depersonalizing the interview, computerized survey administration can help overcome respondents' tendency to skew their answers in the direction of what they believe the interviewer would like to hear. The use of computers in interviews not only diminishes these problems, but provides an opportunity to assess their importance. For example, audio-CASI would make it simple to test the effect of interview pace on the responses given. Indeed, gauging this effect is important, because it is well known that interviewers proceed more hurriedly when they are paid for completed interviews than when paid by the hour.

Researchers can also use computers to avoid the potentially confounding effects of cross-interviewer differences in willingness to repeat or clarify a question in a uniform manner. One bugaboo of survey research is how to make sure respondents understand the question while avoiding tainting interviews with spontaneous (and forbidden) explanations by the interviewer. With audio-CASI, if respondents have the option of saying that they do not understand a question, the computer could offer a standard explanation, or even a set of sequenced explanations or clarifications.

Data accuracy

The cultural, social, and economic context in which a survey is conducted matters in other important ways as well. In particular, what is sensitive in many contexts and cultures may go well beyond sex, crime, and drugs. For example, questions about income, spending, assets, and educational attainment have proven to be highly intrusive in many contexts. Depending on family dynamics and norms of conduct and fairness, individuals may perceive an incentive to hide or understate their financial wherewithal. For example, individuals who divulge their income or assets in the presence of household members not previously privy to such information could face demands for greater contributions toward household expenses. Similarly, disclosure of total expenditures could provoke suspicions about various frowned-upon activities, such as gambling, alcohol and tobacco consumption, and secret support of non–household members.

Even determining a respondent's marital history can be problematic. In the early 1980s the U.S. Bureau of the Census discontinued asking questions about men's marital status and age at first marriage in the quinquennial marriage and fertility history supplement to the Current Population Survey. The data collected were patently inaccurate as judged by reference to vital registration data (*2*). Many men who had been married before but were not married at the time of the survey reported (either directly or by proxy) that they were single rather than widowed, divorced, or separated. Many married men seem to have erroneously reported their current marriages to be their first, apparently to keep their current wives from knowing about their previous marriages. Interestingly, the marriage data for women showed no evidence of similar problems.

Although much research remains to be done on the reasons for intentional misstatement by survey respondents, the problem is well established. Turner *et al.* used audio-CASI to test the validity of responses to

The author is at the Harvard Institute for International Development, Harvard University, Cambridge, MA 02138, USA. E-mail: Dbloom@hiid.harvard.edu

questions believed to be sensitive. As noted earlier, the range of questions falling into this category is larger than just those pertaining to sex, drugs, and violence, and testing this supposition by extending the experiment to a broader range of questions might prove enlightening. An audio-CASI experiment might reveal that responses to questions treated as nonsensitive have been routinely misstated in previous surveys. Such an experiment could also be used to establish cultural differences in the sensitivity of particular questions, and thereby enrich our interpretation of previously collected data and improve the design of future surveys.

Ensuring data quality is perhaps the weakest link in the survey research process. For example, the World Bank's Living Standards Measurement Study (LSMS), a large household survey that the World Bank has conducted in approximately two dozen countries since 1985 and that has been used extensively in academic and policy research, involves little independent confirmation of the validity of the data gathered (*3*). Like most surveys of households and individuals, the LSMS attempts to safeguard data quality by adhering to careful procedures for data collection and coding. High response rates, internal consistency of the data, competent and well-trained interviewers and technical teams, and avoidance of proxy respondents for adults are all indicators of a high-quality survey process; however, these indicators provide little direct evidence that the resulting data are accurate. The Turner *et al.* results provide evidence that is consistent with the possibility that standard approaches to household data collection are seriously flawed. At the same time, they demonstrate that researchers can use randomized experiments in conjunction with new information technologies to assess data quality more broadly and persuasively and to remedy possible deficiencies.

Further explorations

The use of experimental techniques in the design of surveys has a long and successful tradition. Schuman and Presser (*4*), for example, summarized decades of research on carefully designed experiments on the form of questions (for example, open-ended versus closed), their wording (for instance, tone and neutrality), the overall structure of surveys (for example, the order of questions and response options), and the survey context. Their main point is that these factors powerfully influence the quality of survey data. Turner *et al.* further this tradition, but have only scratched the surface of what is possible. Their work suggests a range of follow-up experiments:

First, the Turner *et al.* design is a two-armed randomized experiment that compares self-administered paper questionnaires with audio-CASI surveys. An experiment that included a third arm—the traditional oral survey—could buttress the conclusion that sensitive behaviors are underreported.

Second, the addition of another round of data collection could also yield valuable evidence about the effect of the mode of data collection on data quality. All of the respondents could be re-interviewed after being re-randomized among survey modes. If audio-CASI is truly able to elicit more accurate data, one would expect a higher reported rate of sensitive behavior by individuals surveyed by this method during the second round who took self-administered paper questionnaires during the first round and vice versa. Individuals surveyed by the same method on both rounds would provide an estimate of the composite effect of recall bias, possible new sensitive behaviors between the first and second rounds, and a possible effect of the first interview on the results from the second interview.

Third, audio-CASI could be further tested by an experiment focusing on truly nonsensitive information. Turner *et al.* touch on this issue with respect to heterosexual behavior, but their results may be somewhat inconclusive, as alluded to in an explanatory footnote [table 2, footnote d, in (*1*)].

Conclusions

Randomized experiments offer a powerful methodology for illuminating potential weaknesses in the quality of survey data. Ultimately, however, they cannot decisively validate any particular set of results. This is so even when, as in the article, the new data do seem to correspond to the results from a preponderance of independent studies, because each study is subject to common or idiosyncratic flaws and biases of its own. In general, researchers must understand that respondents have a complex set of motivations for responding as they do, and it is possible that they have some undiscerned motivation for answering questions incorrectly when the questions are delivered via a combination of computer and headphones.

Notwithstanding the promise computers hold for the collection of high-quality survey data, some practical observations from the field may temper social scientists' enthusiasm in this domain. First, these new technologies require careful training of interviewers, not only to ensure that they have mastered the technology themselves but to be certain that they can quickly and effectively show respondents how to use it. This last point cannot be taken for granted, because the respondents for whom this technology may make the most difference sometimes have low levels of literacy and may be unfamiliar with and intimidated by computers. Second, the task of programming a computer to implement skip patterns correctly and to anticipate all of the pathways that might be triggered by particular combinations of responses is far from trivial, and extensive pretesting is required to minimize errors. If mistakes are discovered in the field, the time required to correct them can be large compared to the planned period of survey administration. Such an outcome can wreak havoc with the collection and analysis of data. Third, lest we think that computerization of survey administration will save money, experience to date has often revealed little, if any, savings. And fourth, we must remember that the potential for computers to enhance data quality is still limited by some of the same factors that make the collection of high-quality data difficult in the first place. For example, collecting data in poor, rural homes, as opposed to a controlled, laboratory setting, may affect individuals' responses, even when privacy is unquestioned.

The longstanding issue of how to assess and improve data quality is still with us. The complexity and difficulty of this problem creates a natural temptation among social scientists to neglect it. This problem is aggravated by an underdeveloped tradition in the social sciences of replicating studies using independent datasets and by the fact that the implications of empirical studies are often difficult to test rigorously. Using new technology with experimental methods gives us a valuable tool to approach this issue, especially if the work is done in tandem with psychologists and anthropologists to understand human motivation better and to explore other means of validation, such as intensive follow-up interviews. There is no feasible strategy on the horizon for developing a magic bullet that will solve the problem of survey data quality. Instead, the most promising approach will probably involve multiple disciplines and methods that will provide us with a set of complementary indicators of data quality and guidelines for improving data collection, with new information technologies and further use of randomized experiments figuring prominently in these efforts.

References

1. C. F. Turner *et al.*, *Science* **280**, 867 (1998).
2. A. Pendleton, J. McCarthy, A. Cherlin, *Assessing the Quality of Retrospective Marriage Histories: The June 1980 Current Population Survey* (unpublished manuscript, 1984).
3. M. E. Grosh and P. Glewwe, *A Guide to Living Standards Measurement Study Surveys and Their Data Sets*, LSMS Working Paper No. 120 (World Bank, Washington, DC, 1995).
4. H. Schuman and S. Presser, *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context* (Academic Press, San Diego, CA, 1981).

5. I thank L. Rosenberg, N. Bennett, T. Brown, T. Croft, A. Gallup-Black, M. O'Connell, and R. Shapiro for comments and discussions.