

ARTICLES

Comparability of Measurement in Public Opinion Polls

George F. Bishop*, Stephen T. Mockabee†

Keywords: survey practice

<https://doi.org/10.29115/SP-2011-0031>

Survey Practice

Vol. 4, Issue 6, 2011

Comparability of Measurement in Public Opinion Polls

A *cardinal assumption* we make in asking any survey question is that it should mean essentially the same thing to all respondents. Experienced survey research practitioner Floyd (1995, p. 84; cf. Belson 1981) has expressed this principle in his recommendations for improving question wording: "...A survey question should be worded so that every respondent is answering the *same question* [emphasis added]." Furthermore, as survey methodologist Groves (1989, p. 450) summarized the issue: "Although the language of the survey questions can be standardized, there is no guarantee that the meaning assigned to the questions is constant over respondents." This becomes critical because "A fundamental tenet of scientific measurement is that the measuring device is standardized over different objects being measured (Groves 1989, p. 449)." Indeed, without such standardization how would we determine if our measurements are reliable and valid? Over a quarter of a century ago, political scientist Brady (1985, p. 269) reminded us that the "lack of interpersonal comparability of survey responses" was a greatly neglected and "...serious difficulty... largely ignored by social scientists." Nearly 20 years later, King and colleagues (2004) reminded us again of the potentially "devastating consequences" of ignoring the incomparable responses given to most survey questions, while calling for the increased use of anchoring vignettes to enhance the comparability of measurements. Their wise counsel, and that of Belson, Brady, Groves, and Fowler, among others, has gone largely unheeded – at our paradigmatic peril.

A corollary to the *comparability principle* is that a survey question should mean the same thing to respondents at time two that it did at time one – what we call the *consistency corollary*. As Norman Nie and colleagues (1979, p. 11) posed the problem in *The Changing American Voter*, "Even if the same question is asked at two different points in time, is it really the *same question* [emphasis added]? The fact that times change may mean that the meaning of the question undergoes change." The real world of politics evolves continuously and with it

* **Institution:** University of Cincinnati

† **Institution:** University of Cincinnati

the meaning of the words we use to describe it in survey questions – words such as “liberal” and “conservative” to take two conspicuous examples. A question asked about the “big government” in Washington, D.C. in the 1950s may not, as Nie and his associates (1979, Ch. 1; cf. Fee 1981) noted, have meant the same thing when re-asked in the turbulent 1960s – not to mention what it would mean today because of the ever-changing context of American politics (e.g., contemporary Tea Party and Occupy Wall St. movements). So the meaning-and-interpretation of survey questions about subjective psychological phenomena thus becomes historically contingent and inextricably bound up with the social, economic and political context at the time of measurement. As a consequence, if basic assumptions about the *invariance-of-meaning* cannot be made, then valid comparisons across survey respondents and over time become enormously difficult, if not theoretically and technically impossible.

Comparability of meaning matters a lot, whether we want to admit it or not. For this, after all, is why we care so much whenever someone tinkers with the wording or context of a question in a time series or whenever we compare the conflicting results of polls conducted by different survey organizations. Researchers realize, of course, that differences in wording or context can significantly alter the *meaning-and-interpretation* of a given question, thus destroying temporal, interpersonal, or inter-organizational comparability. As Page and Shapiro (1992, p. 39) aptly put it: “...question wording matters. What may seem to be very slight shifts in wording... can alter people’s interpretations of what a question means, and thereby alter their responses even while their opinions stay constant.”

How right they are. But as a discipline we still have a real blind spot on this score. Although most public opinion researchers have certainly become sensitive to the effects of variations in question wording and context, they have, with few exceptions, been much less attentive – if not oblivious – to how the meaning-and-interpretation of survey questions can vary across respondents and over time even when the wording and context of the question itself remains identical. How well, for example, do we truly understand this source of measurement error in the standard presidential approval question? What, if anything, do we know about how respondents interpret the question: “*Do you approve or disapprove of the way Barack Obama is handling his job as president?*” Or how he’s handling *the economy*? What does “*handling his job as president*” actually mean to them? What does “the economy” mean? Does it vary across respondents and over time? Would the reader object to changing the wording of the second question to read: “Do you approve or disapprove of the way Barack Obama is dealing with the economic problems facing this country today, such as unemployment or foreclosures of people’s homes?” Of course, because it might change the meaning-and-interpretation of the question. But if respondents interpret the ambiguous phrase, “handling the economy,” in different ways at different times, how is this any different than the effects of a change in wording or context?

Such variations in meaning, we contend, create a survey measurement artifact that is functionally equivalent to changes in question form, wording and context (cf. Schuman 2008, ch. 3). It also represents, in our judgment, one of the most serious threats to the validity of survey measurement, especially when we try to interpret the causal meaning of changes in subjective political and social indicators over time. Moreover, it raises the specter of an epistemological quagmire – whether we can ever ask the same survey question twice, as its psychological meaning-and-interpretation is continuously flowing with the passage of time. It's a *Heraclitan* thing: “Nothing endures but change.”

Yet the discipline remains deeply reluctant to come to grips with the implications of this semantic tsunami. Page and Shapiro (1992, Chs. 1-2), for example, were certainly aware of this measurement error problem in making their case for a collectively rational public. But Page (2007, p. 38; cf. Bishop 2008) gives it but a passing mention in his review of Bishop's book on *The Illusion of Public Opinion*, essentially ignoring the analysis of how events like those of September 11, 2001 can alter the meaning-and-interpretation of survey questions on presidential approval, trust in government, and the like. To be sure, Page and Shapiro take great pains to make sure they are comparing responses over time only to identically worded questions (pp. 30–31, 39) because they realize that their analysis of trends in public policy preferences over time would otherwise become invalidated by question wording artifacts.

If only it were all just a matter of random error canceling out in the aggregate. As Bishop (2005, Chs. 5-6) has argued, in analyzing trends in public opinion indicators pollsters are often comparing apples with oranges, even when dealing with identically worded questions. Real-world events, as they get covered in mass media, can alter the interpretation of what an identically worded question means and thereby change the trend lines, sometimes rather suddenly as in the aftermath of the events of September 11, 2001, and sometimes much more subtly over time as in the case of the American National Election Study question (1956–1973) about “keeping soldiers overseas where they can help countries that are against communism” (Bishop 2005, p. 116). Page and Shapiro (1992) write these fluctuations off as “shifting referents’... in which the wording stayed the same but referred to a shifting reality...yet such fluctuations do not necessarily signify any real changes in collective policy preferences” (pp. 58–59). If only this temporal comparability issue could be so readily handled by such verbal gymnastics.

Couldn't the same thing be said about changing the wording or the order and context in which a survey question is asked? That too just shifts the “referents” for respondents, referring them to a different reality, which in turn alters their interpretations of what a question means and thereby destroys the comparability of measurement. But temporal incomparability, we maintain, is temporal incomparability irrespective of how it is generated: by changes in wording, by changes in question order and context, or by changes in the

meaning of identically worded and sequenced questions. They are all functionally equivalent. Measurement error is measurement error, albeit in a different guise. However, the lack of temporal comparability of meaning in a time series is much harder to disentangle and control than most other forms of error we routinely contend with. Furthermore, it is a form of measurement error that is perilous to the paradigm of public opinion and survey research because this type of systematic error is not random and, therefore, will not cancel out in the aggregate. So the miracle-of-aggregation assumption that underpins the foundation of much contemporary macro-modeling becomes untenable, and with it collective models of how public opinion and political attitudes change over time. So too do the trends in political attitudes and public opinion reported routinely in the mass media become vulnerable to the same source of measurement error: shifts in question meaning that are subject to misinterpretation by pollsters, journalists, pundits. In this fundamental semantic sense, common measures of public opinion become potentially misleading, rather than leading, indicators of the state of the nation.

Schuman and Scott (1987) have argued that the only practical solution to the myriad problems of measuring public opinion “requires giving up the hope that a question, or even a set of questions, can be used to assess preferences in an absolute sense...” The solution must rely “instead on describing changes in responses over time and differences across social categories (cited in Moore 1992, p. 347).” Schuman and Scott’s solution of comparing opinion changes over time and across social categories (e.g., racial groups) works only if we can assume that the meaning-and-interpretation of the question remains invariant. But this is the very assumption we are calling into question.

So what is the survey practitioner to do about all this? King and his colleagues (2004), as already noted, have called for increasing the use of anchoring vignettes to enhance the comparability of measurements. Schuman (2008, ch. 3; cf. Bishop 2005, ch. 9) has made a persuasive case for inserting “random probes” with open-ended questions to get at possible differences in respondents’ frames of reference in answering closed-ended questions. This would enable survey investigators to detect shifts in the meaning of questions over time as well as variations in frames of reference across socio-demographic groups.

In designing new survey questions, pollsters should take full advantage of the cognitive interviewing methods developed by Willis (2005) among many others (e.g., think-aloud and retrospective probing). This would do much to ensure that survey questions are worded in a way that means essentially the same thing to every respondent as well as to the investigator who designed it – comparability in the full sense.

But what to do about the comparability of meaning in all the existing survey questions and time series, such as presidential approval, trust in government, and the like? Inserting random probes into such series – ASAP – would get

the semantic detection indicator rolling so that we can begin to systematically monitor how the meaning-and-interpretation of our questions is evolving over time in response to media coverage of events as well as to socio-cultural, economic, and political changes.

As to the meaning of survey questions that have been previously asked, the best we may be able to do is analyze how responses to such items correlate with related items across social categories and over time. Responses to the standard party identification question have become increasingly correlated with responses to the liberal-conservative identification item over time, in both the NORC GSS and the American National Election Studies, suggesting that the subjective interpretations of “Democrat” and “Republican,” for example, have steadily shifted in response to the polarizing political climate of the last couple decades (Bishop 2005, ch. 6). Much the same correlational tracking can be done for many other public opinion indicators, such as the association between the president’s overall approval rating and more specific questions on the economy, foreign affairs, health care, and the like. Bivariate (and multivariate) monitoring of these trends can do much to complement the typical poll reports of mere univariate tracking, thereby revealing much more about the interconnected “meaning” of public opinion across groups and over time. So we urge a realistic view that achieving comparability is more challenging than heretofore suspected, but not a nihilistic view that all is lost.

REFERENCES

- Belson, W.A. 1981. *The Design and Understanding of Survey Questions*. Aldershot, England: Gower Publishing Co.
- Bishop, G.F. 2005. *The Illusion of Public Opinion: Fact and Artifact in American Public Opinion Polls*. Lanham, MD: Rowman & Littlefield.
- . 2008. “Rational Public Opinion or Its Manufacture?” *Critical Review* 20: 141–57.
- Brady, H.E. 1985. “The Perils of Survey Research: Inter-Personally Incomparable Responses.” *Political Methodology* 11: 269–90.
- Fee, J. 1981. “Symbols in Survey Questions: Solving the Problem of Multiple Word Meanings.” *Political Methodology* 7: 71–95.
- Fowler, F.J. 1995. *Improving Survey Questions*. Sage Publications.
- Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.
- King, G., C.J.L. Murray, J. Salomon, and A. Tandon. 2004. “Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research.” *American Political Science Review* 98: 191–207.
- Moore, D.W. 1992. *The Superpollsters*. New York: Four Walls, Eight Windows.
- Page, B.I. 2007. “Is Public Opinion an Illusion?” *Critical Review* 19: 35–45.
- Page, B.I., and R.Y. Shapiro. 1992. *The Rational Public*. University of Chicago Press.
- Schuman, H. 2008. *Method and Meaning in Polls and Surveys*. Cambridge: Harvard University Press.
- Schuman, H., and J. Scott. 1987. “Problems in the Use of Survey Questions to Measure Public Opinion.” *Science* 236: 957–59.
- Willis, G.B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications.