

Survey Measurement for when ‘You Know it When You See it’ and Others See it Differently

Gary King, Christopher Murray, Joshua Salomon, and Ajay Tandon.

[Abstract]

We address two long-standing problems in survey research. The first is how to measure concepts that we know how to define well only with reference to examples --- freedom, political efficacy, pornography, health, etc. The normal advice methodologists give when hearing “you know it when you see it” is to tell the user to come up with a better, more precise theory, and then measurement will be straightforward. This is clearly the right advice, but it leads to a well-known problem, since more highly concrete questions about concepts like these may produce reliable measurements that are not any more valid. The second problem we address occurs because survey respondents in different parts of the world often interpret the same questions in different ways. For example, the self-reported health status of Ethiopians is very similar to that of residents of Denmark, although every objective indicator indicates that the actual health of the populations of the two countries differ substantially. Much of the literature on this problem has focused on writing better survey questions, but despite half a century of efforts it is generally recognized that many of the most important survey instruments still have this problem.

We have designed an approach to survey instrumentation along with a statistical model that together seem to at least partially ameliorate both problems. The idea is that in addition to trying to glean a more precise version of the concept underlying all examples one can think of, we use the examples in the survey questions. Thus, survey respondents are asked for a self-assessment and for an assessment of several (usually 5-7) hypothetical persons described by written vignettes. For example, two of the vignettes for political efficacy are:

[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.

[Moses] lacks clean drinking water. He would like to change this, but he can't vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

We then ask for a self-assessment in (nearly) the same language as we ask for the respondent's assessment of Jane and Moses: “How much say [does ‘name’ / do you] have in getting the government to address issues that interest [him / her / you]?” For both the self-assessment and vignette questions, respondents are given the same set of ordinal categories in which to respond, for example “(A) Unlimited say, (B) A lot of say, (C) Some say, (D) Little say, (E) No say at all.”

The answers to the vignettes by different people arguably provide a fixed and hence interpersonally comparable measurement scale, and so we are able to use it to adjust the answers to the self-assessment questions. Under our model, for each person and question there exists (continuous and unobserved) *actual*, (continuous and unobserved) *perceived*, and (ordered categorical and observed) *reported* levels of efficacy (or the other concepts discussed). People are assumed to have unbiased, but noisy perceptions of their actual health, but when they turn the perceived value of their health status into a reported health category, different types of people use systematically different threshold values --- hence rendering their raw survey responses incomparable. Although the thresholds differ across people, we assume that the thresholds each person uses are the same for the self-assessment question as for their evaluation of the people described in the vignettes. This enables us to correct for interpersonal differences in survey responses due to measured differences among people and their cultures and languages.

Ideally, the vignette assessments are asked of all respondents, to provide a common metric. However, since the variation in the thresholds across people is modeled as a function of explanatory variables, it is also possible under the model to ask the two types of questions of independent samples under certain conditions.

Ideally also, each self-assignment question would have a corresponding set of vignettes with identical ordinal categorical responses to improve interpersonal and intercultural comparisons, but given the costs of survey administration, this will usually be infeasible. Nevertheless, the model posits the set of self-assignment questions as forming a single factor analysis-type model. Thus, using the vignettes to correct one (or more) of these questions with corresponding vignettes still links in the remaining self-assignment questions, and it therefore also corrects for interpersonal incomparability in all. Having multiple self-assessment questions with different categories for answers is also useful to sharpen precision at different parts of the underlying continuous scale.

Statistically, the model we develop is a combination of a random effects ordinal probit model with cut-points that vary systematically over people, parametrically linked to a factor analysis-like model when multiple self-assessment questions are available. We have implemented the entire model without MCMC technology, and so it is computationally efficient, although we draw on Bayesian technology and simulation to compute some quantities of interest.

We have written survey questions with corresponding vignettes for political freedom, political efficacy, responsiveness of the political system in some areas of policy, and seven separate domains of health (mobility, vision, etc.). We have completed approximately seventy surveys with subsets of these questions covering about sixty countries. The full battery of questions will also be used in the World Health Survey, which is intended to be the first global public opinion poll and is presently in the field in about eighty countries. We show how our model improves measurements by comparison to known or more objectively measured quantities in many countries.

We also briefly report on similar efforts to use vignettes and the model we present here to improve measures of income across countries or over time when the “market basket” of goods typically purchased cannot be standardized or exchange rates cannot be fully equalized, and in some other areas. We speculate that a similar protocol could be of use for a variety of measurement tasks in political science and other fields.