

Anchoring Vignettes:  
Frequently Asked Questions

July 19, 2009

# 1 Introduction

This document gives some frequently asked questions about using vignettes to anchor and adjust survey question self-assessments.

We assume that readers of this FAQ have previously read the paper by King, Murray, Salomon, and Tandon (2002).

## 2 Survey Instrumentation

### Why use anchoring vignettes?

1. The act of measurement involves comparing an object to some standard of measurement. The standard is sometimes called an anchor or a gold standard. Without this standard, measurement will be invalid or meaningless. For survey data, anchors can be external, as when voter turnout reports are validated with information from public records, or when health self-assessments are compared to direct medical tests. Anchoring vignettes provide a comparatively inexpensive way of creating an anchor within the survey context itself. The idea is to compare respondents' self-assessments to the respondents assessments of hypothetical people described in short vignettes that have known characteristics, and to use the latter to adjust the former.
2. By describing levels on a particular concept that are fixed across respondents, vignettes provide scale anchors that enable interpersonal comparisons. In combination with the chopit statistical model or our nonparametric alternative, anchoring vignettes can be used to adjust for differences in the way individuals use ordinal response categories.
3. Anchoring vignettes also may be used as a bridge between different items on the same domain. For example, having vignette ratings on two different items relating to mobility allows for comparisons of the response category cutpoints on the two items and for both items to contribute to the estimation of respondents' mobility levels. Related items in different survey instruments may also be linked through the use of vignettes.
4. The technique is also very useful for measuring complicated concepts that are hard to define fully in theory, but can be defined with reference to examples.

**What are the requirements for the use of anchoring vignettes?** We define two key requirements for the use of vignettes as:

1. *Response consistency*, which requires that each individual uses the response categories for the self-assessment question in approximately the same way that he or she uses them to evaluate hypothetical scenarios in the vignettes. Different individuals may use the categories in different ways.
2. *Vignette equivalence*, which required that the domain levels represented in each vignette are understood in the same way on average by all respondents. Of course, even when respondents understand vignettes in the same way, different respondents may use the categorical responses in different ways in answering questions about the vignettes. Each respondent can still apply different types of DIF.

**What empirical evidence do you have that it works?** Section 6.3 of King, Murray, Salomon, and Tandon (2002). We measured visual acuity using the Snellen “tumbling E” eye chart test and compared it to raw survey self-assessments. The Snellen test ranked the Slovaks as having substantially better eyesight than the Chinese, but the raw self-assessments indicated exactly the reverse. We then applied our anchoring vignette correction and drew the same conclusion as the Snellen test. In this situation, we’d still probably prefer the Snellen test to be used rather than the anchoring vignettes, but it is substantially more expensive to administer (especially when done with proper quality controls) and still error prone. That would seem to provide a clear role for the vignettes.

Section 6.2 of the same paper is on political efficacy (how much say do you have in getting the government to address issues that interest you?) in China and Mexico. There’s no direct physical test of political efficacy, but the correct ranking of the two countries could hardly be more obvious from well known external evidence; there too, the self-assessments get the ranking of the countries wrong, but the vignette correction gets it right. For measuring political efficacy for other countries and to study variation within these countries, no feasible measurement strategy exists other than surveys, and the evidence here too indicates that anchoring vignettes improve the self-assessments.

**How many anchoring vignettes should I ask for each concept I want to measure?** Here are the considerations:

1. Chopit, our parametric statistical model, requires, at a minimum, only one vignette (with two or more response categories). (The logic as to why this is sufficient is the same as that in using logistic regression with a dichotomous dependent variable to estimate a continuous probability.) Our nonparametric method can also work with as few as one vignette. In practice, however, since it is sufficiently difficult to write survey questions, we recommend multiple vignettes. This follows the same advice that survey researchers have given in measuring any concept. The use of multiple vignettes would also be required for certain extensions to the standard chopit model such as the addition of random effects in the threshold equations.
2. More important than how many vignettes are asked is designing vignettes that provide discriminatory power. Thus, the best anchoring vignettes are those which are equally spaced through the distribution of self-assessment answers. (For example, asking how mobile a person is who can run 500 miles in a day is obviously of no use in assessing mobility.) The statistical procedures are most powerful when a vignette is asked near to (and preferably on each side of) each respondent’s self-assessment answer. The implication is that the more diverse your respondents in terms of their actual levels and their threshold variations, the more vignettes should be asked.
3. In our research with WHO, we have usually used 5–7, or sometimes as many as 12, vignettes, but our applications involve a large fraction of the world’s population. Surveys of less diverse populations, such as within a single culture, may be possible to do with many fewer vignettes. When possible, we recommend asking more vignettes during the pretest, and then studying how much information is lost by examining the stability of the  $\gamma$  parameters when dropping subsets of vignettes. Monte Carlo experiments can also be helpful.
4. We’re still doing research on the subject but our current (optimistic) guess is that four vignettes asked of 1/4 of the respondents each may be sufficient when you know where the respondents self-assessments roughly are, and you have good covariates to predict the thresholds. (If so, this would add the equivalent of only one item in terms of time on the survey per self-assessment question and so would not be very expensive to administer.)

5. Roughly speaking, the amount of information the data provide about the actual levels increases at most by  $2J + 1$  in the number of vignettes  $J$ . This maximum speed is achieved when answers to the vignettes are most equally spaced through the distribution of self-assessment answers.
6. If you are interested in having higher resolution in measurement at some point in the scale (such as the bottom), then it pays to include more vignettes in this region.

**Should I ask vignettes questions of all people or a random subset?**

1. Chopit works well even when vignettes are asked of only a random subset of respondents, but chopit is only effective if your survey includes variables ( $V_i$ ) that can predict the threshold variation across respondents. Chopit will be more effective when  $\gamma_1$  (the effect of  $V$  on the thresholds) is stable across respondents.
2. Our nonparametric method only works for respondents who have been asked both vignettes and self-assessments. It can be applied to the subset that have been asked both, or both could be asked of all respondents. The advantage of this method is that variables  $V_i$  that are necessary to predict the thresholds in chopit are not needed here. So if you are sure you don't have  $V_i$ , ask vignettes of all respondents.
3. These decisions can be made best by evaluating actual pretest data.
4. Adaptive survey methods might be highly efficient here, whereby respondents are offered new vignettes depending on their previous vignette answers. In this situation, a CATI system could locate vignettes to the left and right of a respondent's self-assessment by bisecting the continuum using a small subset of a large list of a priori rank ordered list of vignettes. Each respondent would then be asked a relatively small number of vignettes. Be aware, however, that this strategy would require a new statistical method be developed.

**How much expense will anchoring vignettes add?** There are several sources of additional costs:

1. Survey administration time: Obviously, vignettes will take up more time on the survey. This cost can be ameliorated to a degree because our model allows vignettes to be asked of randomly selected subsets of respondents, if you have variables capable of predicting variation in respondents' thresholds. If not, then you can use the nonparametric model, which only applies for respondents who are asked both vignettes and self-assessment questions.
2. Translation costs: Three vignettes asked of one third of the respondents each will add the equivalent of only one additional item in expense, but it will also add three questions to translate. Some of these costs can be ameliorated by using vignettes and self-assessment questions asked in other surveys, such as those in the World Health Survey.
3. CATI costs: Asking questions of random subsets makes computer assisted interviewing techniques helpful.
4. Costs of anchoring vignettes can be held down by choosing them appropriately. For example, choosing two vignettes that are very close to each other will provide repetitive information and thus will be wasteful. Similarly, vignettes that are too extreme can provide little or no information.

5. The costs of adding anchoring vignettes to surveys should be weighed in the context of the potential benefits of the approach and the costs of not adopting the procedure. Anchoring vignettes provide the only currently feasible method of testing for DIF, and a good way to correct for it. If you are reasonably sure you have no DIF, then anchoring vignettes will at least provide you the opportunity to verify this hypothesis empirically.

**When do anchoring vignettes make the most difference?** Any or all of these items will improve the efficacy of the approach.

1. Use highly concrete vignettes. The technique makes the most difference for concepts where self-assessment questions are unavoidably vague but vignettes can be concrete.
2. Design vignettes to be roughly equally spaced through the distribution of self-assessment answers. Those too close to each other will provide repetitive information; those too extreme, will provide little or no information.
3. Carefully pretest the survey instrument, analyze the data with our methods and diagnostics to verify that respondents believe the order of the vignettes is as you intended, and remove vignettes that statistical analyses with chopit indicate are too variable incorrectly ordered or provide little information.
4. Make sure the vignettes are tapping only a single unidimensional concept. We find that the process of writing anchoring vignettes often reveals new concepts or dimensions better than writing self-assessment questions alone. Discovering new dimensions makes it possible to narrow the current concept, hence making it more concrete, and possibly to add another self-assessment question and corresponding vignettes for the new dimension.
5. Ask vignettes for every self-assessment question if possible, although our model allows vignettes to be asked that correspond to only one of the self-assessment questions, and to still use the information in the others if all the self-assessment questions are measuring the same concept.
6. Include variables in the survey that will help predict the threshold values. The better the information content in these variables, the better problems with DIF can be detected and corrected with chopit.
7. If you are unable to find variables that can predict threshold variation, then the nonparametric version of the model can be used to correct DIF, but it only works for those respondents who have both self-assessments and vignette answers.
8. Use chopit with a random effect and our conditional predictive method. This tends to work considerably better than unconditional predictions, especially when good variables to predict the thresholds are not available.
9. If you ask all the respondents both the self-assessment and all the vignette questions, you can use chopit with a random effect and then condition on both responses, which can improve the efficacy of the approach even further.
10. Follow all the usual rules and advice given by survey researchers over the last half century. That is, be careful of question wording, question order, accurate translation of the meaning of different items, sampling design, interview length, social background of the interviewer and respondent, etc.

### **In what order should vignettes and self-assessment questions be asked?**

1. We recommend asking the self-assessment early in the survey and the vignettes some time later.
2. The vignettes should *not* be ordered in the survey according to your understanding of their actual value. One problem with this approach is that respondents will tend to try to make their answers consistent over the set of responses, or may use simple heuristics such as placing one vignette in each response category. Both of these outcomes would compromise the requirement of response consistency between vignette ratings and self-assessments. Further, respondents may have different abilities to remember all the previous vignettes, and those who pick the wrong or an unusual value for the first vignette may feel locked in for the rest. The result will be essentially constant responses for the rest that do not discriminate well and so provide little information.
3. We find that vignettes are best presented to the respondent in randomized or mixed order. In addition, if you have more than one set of vignettes, it is helpful to shuffle the two sets together. Since a separate question follows each vignette, this does not cause respondents to have any additional trouble in understanding the survey instrument.
4. Why not ask the self-assessment question after the vignettes? This would lead to an undesirable priming effect, and a different one depending on the order in which the vignettes were presented. Although we might like respondents to read, internalize, and remember the set of vignettes prior to being asked the self-assessment question, this is infeasible for most respondents.

**How is this strategy affected by the finding in social-psychology that assessments of one-self and others differ?** Assessments of others differ from self-assessments because respondents typically have less information about others. Just as when asked for a self-assessment about a behavior that is not easily retrieved from memory, the respondent typically follows an estimation strategy that leads to differing responses. Although vignettes describe someone other than the respondent, all the information necessary to evaluate this other person is in the vignette. By the “maxim of manner” (Clark and Schober, 1992: 27) the respondent will indeed assume that the researcher has provided all the necessary information, and the respondent need not resort to an estimation strategy.

**Can we avoid DIF by using a panel design, without vignettes?** A two (or more) wave panel study enables one to estimate the effects of explanatory variables that vary over time, without using anchoring vignettes. The problem is that a differencing design is required, meaning that the absolute level of the variable being measured (health, efficacy, etc.) cannot be estimated. Nor can the effects of explanatory variables that are fixed over time (such as sex, education level at first interview, etc.) or change predictably over time (such as age).

### **What issues should I consider when writing vignettes?**

1. Vignettes should be written so that people in different cultures understand them as similarly as possible. Translation is of course essential, as is cognitive debriefing during pretests.
2. We find that concrete vignettes that describe specific people and situations are best able to provide constant anchors, although this will not always apply.

3. Each set of vignettes corresponding to a single self-assessment question should tap a single unidimensional concept. The process of writing vignettes is like the process of testing a theory, in that data (or the examples in the vignettes) tend to focus the mind. As such, the process of writing vignettes tends to have important effects on the concepts themselves. New dimensions are discovered, and the features corresponding to them peeled off, making the original set of vignettes more concrete. And of course sometimes new vignettes and a new self-assessment question are added to measure the new dimension.
4. Be careful of the details. Sex, age, and other variables can enter the vignettes by something as simple as the name used or other references. Ask whether these other variables are providing the needed contextual detail for the respondent, in which case they should be retained, or whether they are adding additional unintended dimensions that could confuse the respondent or the analyst.
5. Ideally, only information that is an integral part of the concept being measured should be part of the vignette description. Everything else should be kept implicitly the same as the respondent (so that DIF remains the same for the self-assessment and the vignette questions).

**Should the vignette describe the age, sex, etc., of the hypothetical person? Should it be self-referential?** Vignette answers are a function of both the actual level of the person in the vignette ( $\theta$ , the same for all respondents) and the DIF applied by each respondent (differing over respondents). We can think of these answers as responses to the portions of the vignette text that are, respectively, (1) an integral part of describing  $\theta$  and (2) words used to package these concepts. DIF is generated by the packaging, which human language of course prevents us from eliminating entirely. Fortunately, to meet the assumption of the model, we need not eliminate DIF. We only need to ensure that the DIF each respondent applies in answering the vignette question is the same as the DIF he or she applies in answering the self-assessment question. As such, the goal of writing vignette questions is to keep  $\theta$  accurately described (and distinct from the actual level of the self-assessment,  $\mu$ ), while making the packaging for each vignette close to the description of each respondent so that the DIF will be the same. Normally this is done by excluding age and as much of the other packaging-related information as possible and letting or explicitly encouraging the respondent to think of the vignette as *describing a person like them*, aside from the difference between  $\theta$  (for the vignette) and  $\mu$  (for themselves).

The implied sex of the name of the person described in the vignette is an issue, since ideally this would be the same as the respondent. Thus, if possible, we prefer the names on the vignettes be changed to match the sex of the respondent. When this is impossible or too expensive, using gender neutral names (Lee, Pat, Terry, Kelly, Leslie, Hillary, Bobby, Chris, etc.) or, in some languages, initials (G.K., T.R., B.C., etc.) may be reasonable substitutes.

In principle, we might think about going another step and writing the packaging to reference the respondent explicitly (e.g., “Suppose *you* were paralyzed from the neck down...”). Unfortunately, self-referential vignettes ask the respondent to do research for us in constructing the counterfactual, which in many areas does not work well (e.g., it is similarly not a good practice to ask the respondent for the causal effect of education on his or her income; a better strategy is to ask for education and income and to leave it to the researcher to estimate the causal effect). Asking a respondent to construct a counterfactual, by holding constant some aspects of themselves and changing others may be outside the experiences and beyond the capabilities of many people not trained as social scientists.

In addition, respondents in many cultures seem to respond superstitiously or overly optimistically to counterfactual situations where bad things happen to them, and they give answers that are more extreme than we would expect. In our experience, response rates and test-retest reliability

also tend to drop when individuals are asked to imagine suffering bad or unpleasant health conditions. Our alternative is for the vignette to describe a different person *like* themselves (which we ensure they understand by using a specific named person in the vignette) rather than some counterfactual version of themselves.

An alternative would be to change the vignettes on the fly in a CATI system so that the packaging is extensive and explicitly equivalent to what we learned about the respondent from previous questions (“Bob is a 26 year old plumber from the South Dakota. . .”), but then the fiction of using a different name becomes more and more tenuous. In our experimentation, and cognitive debriefing, we have found that doing this is almost the same in the respondent’s mind as explicitly describing the respondent in the vignette, and it does not work for the same reason. The best option, therefore, seems to be the approach of describing the people in the vignettes as people like the respondent.

**What has to go wrong for anchoring vignette corrections to bias my results?** Here are several ways to think about this issue:

1. First, for simplicity and since statistical methods can deal with it in fairly straightforward ways, imagine that random perceptual and measurement error were nonexistent. Then what needs to happen for *all* the problems to be fixed is that respondents differ in their interpretation of the vignettes only due to DIF (differential item functioning, or interpersonal incomparability), whereas the responses to the self-assessments must differ due to DIF and the actual values (A) on the concept of interest. In addition, since it is the same person answering both questions, we assume that the nature of the DIF is the same for both. The goal is to estimate A. Self-assessments are misleading by themselves because they give  $DIF + A$ , and vignettes can correct since they give us a measure of DIF, and so the correction is  $(DIF + A) - DIF = A$ . But that’s when everything works perfectly.

Now, the fact that the vignettes are subject to DIF and are interpreted in different ways in different cultures by different people is not a problem in and of itself. In fact, the technique relies on vignettes having DIF too. What would be a problem is if the nature of the DIF differs for the vignettes and the self-assessments. In that case, suppose we have  $DIF_v$  for the vignettes but  $DIF_s$  for the self-assessments, and so our correction would be  $(DIF_s + A) - DIF_v$ , which is the same as  $(DIF_s - DIF_v) + A$ . So the ultimate question is not whether the vignettes have DIF, but rather whether  $(DIF_s - DIF_v)$  or  $DIF_s$  is closer to zero. For precisely the reason that it is the *same* person with the same biases answering both questions,  $(DIF_s - DIF_v)$  will normally be closer to zero than  $DIF_s$ . This is the reason why we find that this technique usually is an improvement over self-assessments alone and why only in rare situations does the correction make things worse.

2. The basic assumption is that a respondent uses the same thresholds to translate their perceptions into a categorical response for their self-assessment as for the vignette assessment. An exception would be Rodney (“I never get any respect”) Dangerfield. If he and others with inferiority complexes rank themselves lower than (even hypothetical) vignettes solely because of this complex, and if the pattern of under-ranking themselves is related to other variables of interest but not controlled for in our analysis, then our approach would be biased. The opposite bias may also be possible, whereby individuals rate themselves more favorably than they do hypotheticals (e.g. because of optimism or wishful thinking). We think, however, that these are extreme situations and that the biases would have to be unrealistically large before an unadjusted approach would do better than our adjustment, even with some degree of bias.

3. See the penultimate section of King, Murray, Salomon, and Tandon for additional disadvantages of the model.

**Why will anchoring vignettes work when we know that putting educational achievement tests on a common scale has not been possible?** The one research area where our approach clearly does not work is educational testing. The difficulty with educational testing is that no matter how carefully you write the common test questions as anchors, test takers will differ in their responses to them according to both DIF *and* their knowledge or achievement. Anchoring vignettes solve the problem in other areas because a respondent's answer is *only* a function of DIF (and estimation variability), and so can be used to adjust the self-assessments. An appropriate anchoring vignette in educational testing would be a test question where all test takers have identical knowledge of the subject being examined, but this is obviously infeasible.

**Is there a simpler way of asking questions so we can avoid any statistical analysis?** Direct measurement, that is without statistical analysis, is preferable when possible. We have tried a variety of simpler strategies in a diverse array of national surveys, but none seem to do remotely as well as anchoring vignettes. For example, we tried asking which of a set of vignettes the respondent is most like, but we found that respondents had a difficult time remembering them all at the same time. Another possibility is to ask if the respondent has a higher or lower level of health/efficacy/etc than the first vignette, and then the second, etc. This is better, but it also does not fully correct for DIF in our experience, and in any event would require assuming that DIF is fully corrected rather than allowing empirical verification.

Another possibility of course is the usual strategy of trying to make the self-assessment question even more concrete. This is always a good strategy, but no matter how obvious and unambiguous a survey question appears to the researcher, respondents always seem to surprise us in their ability to interpret questions in different ways than intended. This surprise of course is not revealed unless researchers debrief respondents in post-interview debriefing sessions. Researchers who are sure that their survey questions have no DIF need to verify this at first with these interviews, but ideally also with at least pretesting with anchoring vignettes.

**Do I need one vignette for each response category?** No, there is no necessary relationship between the two. You may have more vignettes or fewer vignettes than response categories.

**Can I use anchoring vignettes if I don't have variables to predict the thresholds?** Variables that predict thresholds help chopit if they are available. Both chopit and our nonparametric procedure will both work without variables that can predict threshold variation, but both procedures would then require having respondents who are asked both self-assessments and vignettes.

**Doesn't Anchoring Vignettes merely move the problem of coming up with DIF-free survey questions back one level (from self-assessments to vignettes), and so in the end you have the same problem?** No, the goal of survey design under this approach is *not* to design DIF-free vignette questions (which is as difficult or impossible as for self-assessment questions). The approach allows respondents to interpret vignette questions in completely different ways. Instead, the goal of survey design is to write vignette questions that have the *same types of DIF* as the self-assessments, since that provides the necessary information with which we can measure DIF, and with that we can then correct the self-assessments. Since the same respondent will be interpreting *both* the vignettes and the self-assessment questions, the assumptions of the technique are much more likely to be met than having to design DIF-free questions.

**If I have a direct physical measurement, such as a medical test, do I need anchoring vignettes?** The basic process of measurement involves comparing an object under study with some standard. Without the standard, we have no (valid or meaningful) measurement. Anchoring vignettes provide one possible standard, or anchor, to make measurements meaningful. They serve the same purpose as medical tests or other physical measurements when they are available. If you can afford to do the physical tests, and if they are accurate in the area which you are measuring, then you have no need for vignettes as anchors.

For some concepts, direct physical measurement is infeasible. Consider political freedom, for example, where a direct test would involve something absurd like handing respondents a sign denouncing the government, sending them out to the town square, and seeing what happened. Similarly, measuring pain is difficult with direct tests, and measuring many aspects of health system responsiveness directly would also be infeasible. WHO has had great difficulty with medical tests in some areas, in part because the people who administer surveys are good at collecting attitudinal data, and not necessarily good at conducting even simple medical tests in diverse settings. And sending medical personnel to the field can be too expensive, especially on a large scale.

For some dimensions of health, using vignettes to anchor self-assessments may generate less measurement error than using medical or other physical tests. For other dimensions, even if medical tests have less error, they have more error per unit cost of administration. For still others, self-assessments corrected by anchoring vignettes can provide a better inexpensive measurement tool than self-assessments alone. And in other areas, medical tests may be more accurate when they are possible to administer, but noncompliance — when, for example, respondents are asked to bear pain, such as for blood tests, or embarrassment, such as for stool samples or physical examinations, all with more potential benefit to the investigator than the respondent — can be a worse problem than for survey-based measures; although only one component of bias, noncompliance can sometimes take a more accurate medical test and leave its practical application more biased than a survey measure.

**Are universally applicable, culture-independent survey questions possible?** Such a goal is probably not achievable across all domains of inquiry. It is probably not even workable for individual domains in many areas, although it is still important to try. Whether or not universal measurement devices (or universally applicable vignettes) can be invented, we still will often want to compare many aspects of health and other concepts across many different places. Our preference for how to do this in most situations is to get it right in specific contexts, and to build up to more generality when possible by comparing across different small sets of areas in separate studies. If we can get measurement right in each of the villages in one set of studies, for example, that may make it more feasible to compare several of these with different villages in different areas, and ultimately as we improve understanding we might be able to make much broader comparisons.

**Can I use anchoring vignettes to understand why respondents understand survey questions in such different ways?** Yes, once you have anchors in the form of vignettes, you can study the reasons for respondents' different understandings. In the chopit model, the thresholds between the response categories are explained with a set of explanatory variables. We can therefore estimate the effects of these variables on the thresholds. Another way to say this is that the model has multiple systematic components, predicting both the actual values of the concept being measured and the actual thresholds between the response categories, across respondents.

Studies such as these are useful in their own right in understanding how respondents see the world and understand the concepts being studied. They can also be of considerable use in designing better survey questions, especially in conjunction with cognitive debriefing sessions.

### **3 Statistical Analyses with Anchoring Vignettes**

More to come. . . .