Title: Using Anchoring Vignettes to Calibrate Teachers' Self-Assessment of Teaching

Authors and Affiliations:

Kun Yuan, John Engberg, Julia Kaufman, and Laura Hamilton, RAND
Heather Hill, Harvard University
Kristin Umland, University of New Mexico
Daniel McCaffrey, Educational Testing Service

Using Anchoring Vignettes to Calibrate Teachers' Self-Assessment of Teaching

Background

High-quality measures of instructional practice are essential for research and evaluation of innovative instructional policies and programs, as well as for providing feedback to teachers and administrators. Classroom observations are generally considered the "gold standard" for gathering rich information about what teachers do. However, observation protocols can be time-consuming and costly to develop, validate, and implement (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008; Bill & Melinda Gates Foundation, 2012).

Teachers' self-reports through surveys and classroom logs are much more time-efficient and comparatively easy to administer over a large population. However, self-reported results suffer from potential biases due to differences in respondents' understanding of the latent constructs being measured, as well as interpretation and application of the scale used to quantify their practices (Hill, 2005; Chevalier & Fielding, 2011). For example, teachers might have different understanding of what constitutes a cognitively demanding learning task when responding in a survey about the extent to which their students are engaged in such tasks. Furthermore, a high level of an instructional practice as perceived by one teacher may be a lower level as perceived by another. These differences lead to incomparability between teachers' answers and contribute to the lack of validity of teachers' self-reported teaching measures observed in prior research (Mayer, 1999; Hill, 2005; Stecher, Le, Hamilton, Ryan, Robyn, & Lockwood, 2006).

The use of anchoring vignettes in surveys has the potential to diagnose and, in some cases, address the factors that likely lead to inaccuracy in teacher self-reports about their instruction (King, et al, 2004). When using this method, researchers provide an operational definition of an abstract construct to be measured through a hypothetical scenario with detailed descriptions of the cognition and behaviors of individuals similar to the respondents. Researchers change the cognition or behaviors of individuals in the scenario and generate different versions of vignettes to represent different levels of the underlying construct. A group of subject experts rates each vignette using the same scale that respondents will use to rate these vignettes themselves and identify a point on the scale that corresponds to each vignette.

In a typical survey that uses the anchoring vignettes method, respondents are presented with all anchoring vignettes and asked to rate the individual in each vignette as well as themselves on the latent construct of interest. Respondents' ratings on anchoring vignettes are used to calibrate their self-assessments to a common scale so that the adjusted self-assessment results are more comparable across respondents than the raw self-assessment results (Wand, King, & Lau, 2011). This method has been successfully implemented in political science, health, and other fields and found to be useful to calibrate survey self-reports and provide more comparable results across respondents (King, et al., 2004; Grol-Prokopczyk, Freese, & Hauer, 2011; Soest, et al., 2011).

Purpose of Study

In this study, we examined whether using anchoring vignettes in web-based surveys improved the validity of teachers' self-assessments of their mathematics instruction. To investigate validity, we compared correlations between teachers' self-ratings and other measures of teaching including teachers' value-added scores, student surveys, and observation ratings of instruction before and after calibration to examine whether calibration improves the correlation between teachers' self-ratings and other teaching measures.

Significance of Study
This is the first study that uses anchoring vignettes to calibrate teacher self-assessment of teaching. In this study, we will provide a rigorous test of whether response scale differences are partly responsible for the current lack of alignment between teachers' survey self-reports about their instruction and observations of those teachers' instructional practices. If successful, this innovative method could contribute to the development of a new generation of survey instrumentation for assessing mathematics teachers' instruction. Such advancements in survey methods could greatly inform research on the antecedents and effects of instructional practices, as well as provide rapid feedback to school leaders and teachers themselves about their work.

Participants
Data came from 61 mathematics teachers in grades 4-9 participating in the Bill & Melinda Gates Foundation's Measures of Effective Teaching Extension project. The sample was roughly evenly distributed between elementary (grades 4-5) and middle school (grades 6-8) teachers, with only three percent teaching 9th grade. Eighty percent of these teachers were female. About two-thirds were white, one quarter were black, and the remaining teachers were Hispanic. On average, they had about six years of teaching experience in their current districts, and over 40 percent had at least a master's degree.

Data Collection and Analysis
We worked with mathematics education experts to identify six dimensions of mathematics instruction on which to focus in our survey: (1) emphasis on mathematical vocabulary; (2) questioning; (3) emphasis on student effort; (4) use of instructional time; (5) use of cognitively challenging tasks; and (6) remediation. For each dimension, we designed a series of four anchoring vignettes representing four different levels of practice for that dimension.

Teachers completed an online survey within one to two days after video-recording a mathematics lesson. In the survey, teachers rated themselves on these six dimensions for their videotaped lesson. For each dimension, teachers also rated four short anchoring vignettes representing hypothetical classrooms where differing levels of the dimension were present. Finally, they were asked to rank the vignettes, along with their own practices, according to the definitions of each practice provided.

We administered the survey in two waves from January to June 2013, with questions for three randomly-chosen dimensions included in each wave. In the first wave, self-ratings came before the vignette ratings for each dimension; in the second wave, self-ratings came after vignette ratings. Trained raters (two raters per video) scored the videotaped lessons using a rubric that captures the same dimensions as the survey, and they worked together to reconcile any disagreements in their ratings.

We also have a composite teacher performance measure for each teacher, drawn from the original Measures of Effective Teaching project (Bill & Melinda Gates Foundation, 2012). The composite measure is a combined score based on an array of measures, including teachers' value-added scores, results form student surveys, and classroom observations.

Statistical Model

We used both non-parametric and parametric methods to calibrate teachers' self-ratings. The non-parametric calibration method recodes teachers' categorical self-ratings relative to each set of anchoring vignettes. Let y_i be the categorical survey self-assessment for teacher i (i = 1, …, N) and z_{ij} be the categorical survey responses for respondent i on vignette j (j = 1, …, J). For respondents who ranked all four vignettes on each dimension in the same order as the panel of experts ($z_{i,j-1} < z_{ij}$ for all i, j), the calibrated self-rating is

$$c_i = \begin{cases} 1 & if\ y_i < z_{i1} \\ 2 & if\ y_i = z_{i1} \\ 3 & if\ z_{i1} < y_i < z_{i2} \\ \vdots & \vdots \\ 2J+1 & if\ y_i > z_{iJ} \end{cases} \quad (1)$$

Inconsistencies in the ordinal ranking of vignettes are grouped and treated as ties. Respondents with ties in the vignette ratings would receive an interval value for C instead of a scalar value.

The parametric method models the construct with random measurement error and allows the thresholds that turn the unobserved perceived variable into an observed categorical response to vary over individuals as a function of measured explanatory variables.

Let μ_i represents respondent *i's* actual level on the underlying construct to be measured. Assume μ_i is on a continuous, unbounded, and uni-dimensional scale with higher values indicating higher levels on the interested construct. This actual level varies over respondents as a liner function of observed covariates X_i with coefficient β and an independent normal random effect η_i.

$$\mu_i = X_i\beta + \eta_i \qquad \eta_i \sim N(0, \omega^2) \quad (2)$$

The parametric method assumes respondents perceive their actual levels on the interested construct only with random errors. Let Y_{is}^* represent respondent i's unobserved continuous perceived level on the self-rating question.

$$Y_i^* \sim N(\mu_i, \sigma^2) \quad (3)$$

Respondent i answered self-rating question s with K_s ordinal response categories. S/he turns the unobserved perceived level of Y_i^* into the reported category y_i via this observation mechanism:

$$y_i = k \quad if \quad \tau_i^{k-1} < Y_i^* < \tau_i^k \quad (4)$$

The vector of thresholds τ_i (where $\tau_i^0 = -\infty, \tau_i^K = +\infty, \tau_i^{k-1} < \tau_i^k$, k = 1, …, K) varies over respondents as a function of covariates V_i and unknown parameter vectors γ

$$\tau_i^1 = \gamma^1 V_i \quad (5)$$

$$\tau_i^k = \tau_i^{k-1} + \gamma^k V_i \quad k = 2, \ldots, K-1$$

Let θ_j represent the hypothetical person's actual level on the latent construct in vignette j (j = 1, …, J). Z_{ij}^* represents respondent *i's* unobserved perceived level of the hypothetical person in vignette j. Respondent i perceives θ_j with random normal error

$$Z_{ij}^* \sim N(\theta_j, \sigma_j^2) \quad (6)$$

Similar to the self-rating process, respondent *j* turns the unobserved perceived level of Z_{ij}^{*} into the reported category z_{ij} via a similar observation mechanism:

$$z_{ij} = k \quad if \quad \tau_i^{k-1} < Z_i^{*} < \tau_i^{k} \tag{7}$$

The thresholds are determined by the same γ coefficient used for y_i, and the same explanatory variables but with values measured for

$$\tau_{i1}^{1} = \gamma^{1} V_i \tag{8}$$
$$\tau_{i1}^{k} = \tau_{i1}^{k-1} + \gamma^{k} V_i \quad k = 2, \ldots, K-1$$

We implemented the non-parametric and parametric models using the R package anchors (Wand, King, & Lau, 2011). Covariates used in X_i and V_i include teachers' gender, ethnicity, years of teaching experience, and whether the teacher had a master's degree. Then we examined the correlation between teachers' self-ratings and the composite teacher performance measure before and after the calibration.

We also investigated whether teachers rank the vignettes the way we intended; which dimensions teachers and observers rate reliably; and how survey and observation ratings compare.

Findings / Results:
Preliminary findings suggest that anchoring vignettes represent a promising innovation for measuring teachers' instruction through survey self-reports. Specifically, we found:

- Teachers' survey responses that are calibrated through the use of anchoring vignettes have increased variation compared to teachers' raw survey responses, particularly for the cognitive challenge dimension;
- Teachers' calibrated survey responses regarding mathematical vocabulary and cognitively challenging tasks are more strongly correlated with the composite measure of teacher performance compared to raw survey responses;
- If teachers gave their self-rating after rating the vignettes, rather than before, the entire collection of calibrated self-ratings are significantly correlated with the composite performance measure ($p<.05$).

Conclusions:
These findings suggest that anchoring vignettes improve the accuracy of teachers' self-reports, which has implications for how researchers and practitioners can efficiently gather and learn from instructional data.

Limitations of this study include restricted sample size and lack of evidence regarding validity of the measure for specific purposes such as use in a high-stakes evaluation system or use as a source of information to inform professional development.

Future research may refine the practices examined in this study; incorporate practices that we haven not studies as carefully such as classroom climate; identify the specific teacher-student interactions that characterize each level of a given dimension; and study how to use anchoring vignettes or vignettes calibrated survey results for different purposes such as professional development or making high-stakes decisions.

Appendices

Appendix A. References

Chevalier, A. & Fielding, A. (2011). An introduction to anchoring vignettes. *Journal of Royal Statistical Society, 174(3)*, 569-574.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*, Author: Seattle, Washington.

Grol-Prokopczyk, H., Freese, J., & Hauer, R. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior, 52(2)*, 246-261.

Hill, H.C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, 19(3), 447-475.

King, G., C.J.L. Murray, J.A. Salomon, and A. Tandon. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1), 191-207.

Mayer, D.P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, 21(1), 29-45.

Pianta, R.C., J. Belsky, N. Vandergrift, R. Houts, and F.J. Morrison. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, 45(2), 365-397.

Soest, A., Delaney, L, Harmon, C. Kapteyn, A., & Smith, J. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of Royal Statistical Survey*, *174(3),* 575-595.

Stecher, B., V.-N. Le, L. Hamilton, G. Ryan, A. Robyn, and J.R. Lockwood. (2006). Using structured classroom vignettes to measure instructional practices in mathematics. *Educational Evaluation and Policy Analysis*, 28(2), 101-130.

Wand, J., King, G., Lau, O. (2011). Anchors: Software for anchoring vignette data. *Journal of Statistical Software, 42(3)*, 1-25.