

BARBARA AUSTIN

Northern Arizona University

Rubricizing Teaching for Validity

Vexation

Teacher self-reporting is ubiquitous in education research and program evaluation. For example, researchers may interview teachers about their beliefs and practices. Or, teachers could be asked to fill out a Likert scale (e.g., "Agree", "Strongly Agree", etc.) survey about the most common techniques they use in the classroom. Likert surveys are easy to administer and evaluate, only requiring a few marks by the respondent and, at worst, a Scantron machine to score the results. Even easier are online surveys that can be taken by teachers at their leisure. Large amounts of data can be generated relatively painlessly on the part of the teacher and the researcher or program evaluator. These large amounts of data are easy to analyze for reliability. Unfortunately, teacher self-reporting has many shortcomings and flaws that outweigh the ease by which the data accumulate. It is my view that we should avoid treating self-reporting as valid sources of information.

Typically, teacher self-reporting documents include force-choice responses. A major drawback of this type of evaluation is that there is essentially no mechanism for determining behavioral motivation. Teachers may misinterpret or misunderstand items, completely invalidating their responses. Teachers' recall of events and behaviors may be inaccurate or they may misinterpret classroom interactions. They may deliberately misrepresent their feelings or behaviors. For example, in the SALISH study, researchers from 9 teacher education programs studied 69 beginning teachers in their first three years of teaching. The researchers used three instruments to observe teachers: two consisted of self-reporting instruments (a Likert scale survey and an interview protocol) and the third was a coding scheme for classroom observations. The researchers found that although self-reporting of teacher-centered beliefs declined in the first three years, student-centered behaviors also declined while teacher-centered behaviors increased. In other words, observations of classroom practice did not match self-report.

Another problem associated with self-reporting is that through both preservice and inservice training, teachers know what the 'right' answers should be to questions about teaching. In case you want to study up for the next survey or interview, the right answers include: 'constructivist', 'student-centered', 'hands-on', 'high expectations', 'inquiry', 'success for all', 'critical thinkers', and 'reflection'. I suspect there is a correlation between the amount of training teachers receive and their propensity to talk the talk. Unfortunately, this training may not translate into walking the walk. In a recent study of 300 5th grade teachers, researchers from the College of William and Mary found no significant correlation between the student achievement of nationally board certified teachers and non-board-certified teachers. In a second phase of the study 21 board-certified teachers rated higher than non-board-certified teachers on interviews and submitted assignments, both of which are types of self-reporting. However, a group of 16 non-board-certified teachers who were identified by student achievement data, received higher ratings from classroom observations by the researchers. In other words, the teachers whose students had the largest achievement gains were also the ones the researchers rated highest during classroom observations. The board-certified teachers were better at talking about teaching but this superiority of articulation failed to translate into higher student achievement. Thus we see from this example that teacher self-reporting is an unsatisfactory source of information about teaching.

The lack of correlation between National Board certification and student achievement, the results of the SALISH study, and the correlation between observed practice and student achievement imply that teacher self-reporting is problematic. However, from my short experience as a professor, it seems to be the standard method for evaluating effectiveness of teacher training programs. The question is why?

Venture

Much of my graduate school education was paid for by the National Science Foundation, for which my family and I are extremely grateful. A large portion of this work was spent in classrooms evaluating and supporting teacher implementation of strategies from the professional development program I worked with. We introduced teachers to a method of problem-based learning using a web-based shell called a Legacy cycle. Legacy cycles scaffold the problem-solving process experts use to generate new knowledge including steps of brainstorming, researching what others have to say about problem solutions, recalling or learning new knowledge pertinent to the problem solution, testing hypotheses, and presenting findings. Teachers stayed in the program for two years. All of the teachers created Legacy cycles that met the goal of scaffolded problem-solving. However, in implementing the cycles, most of them interacted with students and the content in very traditional ways in being sources of knowledge rather than facilitators of knowledge discovery. They learned how to plan problem-

Rubricizing Teaching for Validity

based lessons but were not able to implement them in a way that was truly consistent with problem-based learning. For example, they told students how or where to find answers to questions rather than letting students discover answers on their own.

Although I love being in classrooms, collecting meaningful data on teacher practice is subjective, tedious, time-consuming and leads to carpal tunnel syndrome while transcribing field notes. Data analysis is best accomplished using programs like *NVivo*, which have steep learning curves, even with training. This creates some genuine tensions: the qualitative data is a step above teacher self-report; it seems much more authentic and likely to approach trustworthiness. But, the cost in terms of time and energy is extraordinary.

As an alternative to the qualitative data I have collected, I have also used observation rubrics to supervise student teachers both when I was a classroom teacher and a university supervisor in graduate school. However, from my short experience, because of the response choices, rubrics do not capture teacher behavior and motivation—two key elements of behavior targeted by reform-based teacher education programs. The rubrics I used had response choices along the lines of ‘emergent’, ‘proficient’, ‘highly proficient’, and ‘exemplary’. Regardless of the content or wording of the items, these responses don’t describe what the teacher was actually doing. For example, a supposedly identifying characteristic might have been “communicates effectively” or “uses inappropriate questioning technique.” Such descriptors give no idea about what the teacher was actually doing.

Because of the generic nature of these rubrics, others who were unfamiliar with that teacher wouldn’t be able to make their own judgment about the validity of the teacher evaluation or draw their own conclusions about the efficacy of the practice. Another problem with many rubrics is that, because of the hierarchical nature of responses, they are easily manipulated by people with an agenda, such as principals, who have a vested interest in the appearance of teacher quality. A related problem that I saw while supervising at the university level were teachers-turned-supervisors who didn’t want to give student teachers the equivalent of a “C” on the rubric, even if that was the level at which students were actually performing.

Because of the problems already mentioned with regard to qualitative data, the alternative to teacher self-reporting will mostly likely have to take the form of a rubric. This instrument would need to be constructed in a way that captures the unique classroom practice of individual teachers. An additional feature is that ‘right’ answers must be opaque to potentially biased observers. For example, rather than creating hierarchical responses of “exemplary”, “proficient”, and “insufficient”, it might be better to generate responses for each item that actually describe the evaluator’s experience while eclipsing ‘right’ answers. Here is one example:

Table 1: Proposed observation rubric

Item	1	2	3	4
1. Instructional delivery	Instruction focuses on minutiae rather than broad concepts	Examples provide tangential support to main topic	Concepts are presented without contextualization	I can’t rate this item because learning objectives are unclear
2. Organization	I would never have learned to tie my shoes if this teacher were my only pathway to learning	Topic sequencing seems intended to lead to broad understanding of discipline rather than finite factual knowledge	The lesson appears to take a ‘textbook path’ through the content	I can’t rate this item because learning objectives are unclear

Another characteristic of a proposed rubric is that instead of capturing the observation holistically (these are often filled out at the conclusion of the entire lesson), the rubric should be constructed to be filled out while the teacher is teaching in order to try to capture the actual experience of the lesson. Consequently, many items may appear multiple times on the rubric.

For my venture I could get on my soapbox and preach to the education community about the futility of expecting to be able to generate accurate knowledge about teacher education when self-reporting is one of the fundamental instruments. The odds of success seem very small. However, I remain hopeful and would like to use *Crossroads* as a forum to find an alternative to teacher self-reporting. Here are some things I would like to know:

- a. What potential and problems are embedded with using descriptive rather than hierarchical rubrics?
- b. How could these types of rubrics be flexibly analyzed and validated?
- c. Where might one begin to generate a groundswell of researchers willing to abandon teacher self-reporting?