NEPC
NATIONAL EDUCATION POLICY CENTER

Review of *Asking Students about Teaching*

Reviewed By

Eric M. Camburn

University of Wisconsin-Madison

November 2012

Summary of Review

Asking Students about Teaching seeks to establish that student surveys provide valid evidence that can be used for evaluation of and feedback for teachers. The report then proceeds to advise practitioners about optimal practices for administering student surveys and using survey information. As the report contends, student surveys are a useful tool in practitioners' and policymakers' toolkits, and the report contains many practical pieces of advice that are sensible and worth putting into practice. But a major limitation of the report is that the claimed relationship between student survey reports and teacher effectiveness is not supported by the evidence provided. A broader limitation of the report is that many of the findings and conclusions are presented too uncritically and without sufficient justification. Developers of the MET project embrace the idea that multiple measures of teaching effectiveness are needed to represent such a complex, multi-faceted phenomenon. In discussing the potential uses of student surveys, however, this report's stance is lopsided, placing too much weight on the strengths of student surveys and not enough weight on their weaknesses. A potential concern is that glib implementation of some of the report's recommendations might result in an unwarranted overconfidence in student survey results.

Kevin Welner
Project Director

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar

This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

Review of *Asking Students About Teaching*

Eric M. Camburn, University of Wisconsin-Madison

I. Introduction

The Bill & Melinda Gates Foundation's "Measures of Effective Teaching" (MET) Project endeavors to identify measures that accurately demonstrate how well teachers help their students learn. A major assumption underlying the project is that "multiple measures" are needed to give a complete picture of teachers' effectiveness. Based on this assumption, MET focuses on how multiple measures of effectiveness should be combined to accurately capture the multiple facets of teacher effectiveness. The MET project is vast, involving 3,000 teachers in six school districts, researchers from five universities, and partnerships with various non-profit organizations and education-related companies.

The report reviewed here, *Asking Students about Teaching: Student Perception Surveys and Their Implementation*,[1] is the third MET project report released by the Gates Foundation. The first report, *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*, was released in December 2010. It examined correlations between student survey responses and value-added scores computed both from state tests and from higher-order tests of conceptual understanding. The study found that the measures are related, but only modestly.[2]

The second report, *Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains*, focused on classroom observation protocols as potential measures of teacher effectiveness. That report found that the observation instruments examined have fairly low reliability and are only weakly correlated with value-added measures.[3]

Asking Students about Teaching has two primary purposes. It first sets out to establish that student surveys provide useful evidence of teaching that can be used for teacher evaluation and feedback for teachers. Much of the evidence used to test this proposition comes from the first two MET reports. After presenting this evidence, the report proceeds with its second purpose of providing advice to practitioners about optimal student survey practices. It's worth noting that *Asking Students about Teaching* is identified as a "Policy and Practice Brief" while the first two MET reports were "Research Papers."

The report provides a good deal of sensible practical advice. For example, it recommends that students' answers on surveys will be more accurate and truthful if they feel their answers won't be seen by their teacher or fellow students. It also suggests that evidence

about a teacher's performance from 30 of her students could carry as much or more weight than evidence from two or three observations by an outside observer. While many such recommendations are logically sound, few are solidly grounded in methodological or empirical literature.

There are a number of serious concerns about using student survey data as a valid indicator of teacher effectiveness. The report minimizes some of the challenges practitioners and policymakers would likely face in practice. This review identifies a number of areas where greater caution should be considered.

II. Findings and Conclusions of the Report

The report includes four main findings. Each finding supports the contention that student surveys provide valid evidence of teaching effectiveness and provide useful feedback for teachers.

1. "…teachers' student survey results are predictive of student achievement gains. Students know an effective classroom when they experience one" (p. 2).
2. "…student surveys produce more consistent results than classroom observations or achievement gain measures" (p. 2). "…the MET project found Tripod[4] to be more reliable than student achievement gains or classroom observations…" (p. 14).
3. "…students who completed the Tripod survey as part of the MET project perceived clear differences among teachers" (p. 5).
4. "The MET project's analysis of Tripod found [the conclusion that the average student survey responses for teachers generally predicted teacher value-added scores] to be true for the sample of teachers it studied" (p. 9-10).

The report also included eight conclusions about optimal practices for using student survey results as evidence of teaching effectiveness.

1. "Survey items need to be clear to the students who respond to them" (p. 11).
2. "If students believe their responses will negatively influence how their teachers treat them, feel about them, or grade them, then they'll respond so as to avoid that happening. … They should be told, in words and actions, that their teachers will not know what individual students say about their classrooms" (p. 12).
3. "[Student survey] systems must be certain about which teacher and class each completed survey belongs to. Part of ensuring this requires making sure students have the right teacher and class in mind when responding." (p. 12).
4. "Both reliability and feedback can be enhanced by including multiple items for each of a survey's constructs" (p. 14).
5. "Even a comparatively reliable measure could be made more so by using bigger samples. …averaging together results from different groups of students for the same

teacher would reduce the effects of any variance due to the make-up of a particular class" (p. 16).

6. Understanding what results from surveys mean "…requires knowing each question within the construct, one's own results for each item, and how those results compare with those of other teachers" (p. 18).

7. "For most people, improvement requires the example and expertise of others. While student surveys can help point to areas for improvement, they can't answer the question 'Now what?'" (p. 19).

8. "Although often ill-defined, collaboration is key to effective implementation of new practices… Every system referenced in this brief has made stakeholder engagement central to its rollout of student surveys…" (p. 21).

III. The Report's Rationale for Its Findings and Conclusions

Asking Students about Teaching reports empirical findings from previous MET reports and related analyses. These findings differed considerably in how thoroughly they were explained and in the extent to which they were grounded in evidence or theory. In every case, however, important details were missing about the methods used and statistical results obtained. Regarding the eight optimal practices for administering student surveys about teacher effectiveness, few of the conclusions were justified with appeals to methodological, theoretical, or empirical literature despite the fact that relevant literature exists in a number of cases.

It is worth noting that *Asking Students about Teaching* is a "Policy and Practice Brief" that is intended for a non-technical audience of practitioners and policymakers. This review takes the position that regardless of intended audience, the scientific foundation supporting the recommendations should be presented or, at least, referenced.

The Report's Four Findings

The report asserts that "student survey results are predictive of student achievement gains" and concludes that "Students know an effective classroom when they experience one" (p. 2). These assertions appear to be based on results from the first MET report, *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*.[5] That report found modest correlations between teacher value-added measures in mathematics and reading and student survey measures. Correlations with mathematics value-added measures disattenuated for measurement error ranged between .3 and .49. Correlations with English Language Arts value-added measures were even lower. The unqualified, strong assertion that "Students know an effective classroom when they experience one" is not warranted by these modest correlations accounting for, at best, less than 25% of the variance. Moreover, neither statistical results nor methodological details about the original analysis are provided in *Asking Students about Teaching*, nor does this new report cite the original report containing the statistical analyses.

The second finding reported in *Asking Students about Teaching*, that the Tripod student survey is "more reliable than student achievement gains or classroom observations" (p. 14), appears to come from the second MET report, *Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains*.[6] Table 16 of that report provides reliability statistics for the Tripod student survey, four classroom observation protocols, and "student achievement gains." The table does indicate

> *The authors' unqualified conclusion that students perceive "clear differences" in their teachers is not warranted.*

that the reliability of the Tripod student survey (.65) is higher than that of all four observation instruments (.20, .40, .42 and .43) and of student achievement gains (.48). Thus, based on the earlier report, the finding appears to be technically correct. It is worth noting, however, that all these reliability statistics are quite low and would warrant caution if these instruments were being used for high-stakes purposes. Again, though, neither statistical results nor methodological details about the original analysis are provided, nor is the second MET report cited.

The third finding, that the Tripod survey measured "clear differences" in how students perceived their teachers, came from an original analysis of Tripod data. Measures of teaching effectiveness were created by calculating for each teacher the proportion of 36 Tripod items their students answered "favorably." Tripod uses five-point rating scales that differ between elementary and secondary survey versions.[7] Two groups of teachers were identified, one at the 25th percentile on this measure and another at the 75th percentile. The authors then compared the two groups on the percentage of students who agreed with seven student survey items (one item apiece from each of the "7C" constructs, as identified below in endnote #4). For all seven items, students whose teachers were at the 75th percentile of overall favorable ratings were substantially more likely to agree than students whose teachers were at the 25th percentile. Nonetheless, the authors' unqualified conclusion that students perceive "clear differences" in their teachers is not warranted. The results merely indicate that students in classrooms in which more students agree with Tripod items overall are more likely to agree with a particular item than students in classrooms with lower overall levels of agreement. Because the two measures being compared come from the same set of items from the same survey administered to the same set of students, the reasoning is circular.

The fourth finding, that students' perceptions of teachers "generally predict" achievement-based measures of teacher effectiveness, was based on an analysis in which teachers' value-added scores were related to the average Tripod survey scores of their students. Separate analyses were conducted for mathematics and English Language Arts achievement. Student ratings were measured by summarizing the results from the 36 "7C" items and then computing teachers' percentile rank on the summary measure. The authors reported that teachers "…in the top 25 percent based on Tripod results had students who were learning the equivalent of about 4.6 months of schooling more in math … than

students of teachers whose survey results were in the bottom 25 percent" (p. 10). In addition, the report includes graphs depicting what appear to be regression lines for the mathematics and English Language Arts analyses. The graph for math value-added scores depicts a strong relationship, while the graph for English Language Arts depicted a weaker, but still positive, relationship.

Based solely on the graphical evidence, the authors' conclusion that student perceptions "generally predict" teacher-value added may be fitting. However, basic statistical information, such as the multiple R, statistical significance of regression coefficients and R-squared statistics, are missing. Such essential information would have helped readers understand the rationale behind the analyses and judge whether the authors' conclusions are warranted. For example, the omission of R-squared statistics raises questions about the authors' conclusions. Low R-squared statistics for these models would mean that student ratings were inaccurately predicting teacher value-added scores; under these conditions the graphs could be misleading. In addition, the value-added estimates used in these analyses were not from the same year the student survey data were collected, and the logic behind this approach was not provided. The metric of the value-added measures was also omitted, thus precluding readers from understanding if the relationships depicted in the graphs were of practical importance.

Note that these are largely critiques of the ways in which evidence in support of student surveys was reported. Perhaps many of the concerns raised here could be addressed by the release of a more complete report that provided greater detail about findings and conclusions.

Student Survey Practice Recommendations

The report makes many sensible recommendations for using student surveys for measuring teaching effectiveness. Unfortunately there is no single body of literature (methodological or conceptual) that can be used to support the report's recommendations for using student surveys for data-driven decision making. With regard to the eight specific recommendations for survey practice, however, some relevant guidance can be found in the survey methods and measurement literatures.[8] Some of the eight recommendations are somewhat consistent with principles found in these literatures, while others are not. The main concern here is simply that the rationales provided were not grounded in, nor did they explicitly reference, any of these literatures or other useful evidence.

IV. The Report's Use of Research Literature

Asking Students about Teaching neither cites nor references any empirical or conceptual research. In fact, even though the report presents results from the first two MET reports, those reports are not cited. As previously discussed, there are multiple literatures that

could have been used. Grounding the report's recommendations in the literature would help policymakers more validly and accurately assess the weight of the report's findings and conclusions.

V. Review of the Report's Methods

A review of the first MET report[9] critiqued the teacher value-added models used to produce the first finding: to wit, student surveys about teachers are predictive of student achievement gains. A review of the second MET report[10] critiqued the classroom observation and teacher value-added methods underlying the second finding that student surveys produce more consistent results than classroom observations or achievement gain measures.

As previously discussed, the analysis producing the result that students "perceived clear differences among teachers" used a problematic approach. It compared the survey responses of students in two kinds of classrooms—those with teachers at the 25th percentile of a summary score from the Tripod survey and those at the 75th percentile. Student responses to seven items (one from each of the "7C" constructs) were reported using simple bar graphs. As discussed earlier, the relationship between the two measures being compared in this analysis appears to be tautological. A second problem with this analysis is that defining teacher groups as those *exactly* at the 25th and 75th percentiles of "favorable student responses" on the Tripod survey would appear to limit the analysis to only two points on the overall percentile distribution. The scores on the other 98 percentile points are ignored. The issue is that the 25th and 75th percentile points are but two scores on the overall percentile distribution. Singling out teachers with only these two scores on the overall distribution of Tripod scores would have limited this analysis to a small, idiosyncratic group of teachers.[11]

The finding that Tripod student survey results were "generally predictive" of teacher value-added scores appeared to have been based on regression analyses in which students' ratings of their teachers on the Tripod survey were used to predict teachers' value-added scores in mathematics and English Language Arts a year later. A few descriptive results, apparently derived from regression models, were presented in the text, and the results of the regressions are depicted in two graphs in which the results were plotted. As discussed earlier, the authors omitted the essential details about these analyses—the slope estimates, the metric of the dependent variables (teacher value-added measures), statistical significance of estimates and "goodness of fit" statistics. Moreover, the regression lines, as presented in the report, were not straight but curved. Relationships involving curved lines can be complicated to explain. The meaning of the finding that teachers ". . . identified as being in the top 25 percent on Tripod results had students who were learning the equivalent of about 4.6 months of schooling more in math. . ." is difficult to understand given the kind of complex, curved relationship between student perceptions and math achievement depicted in Figure 2 on page 10 of the report. In general, the limited details

offered for the analyses do not provide sufficient support for a claim that student survey reports provide reasonable evidence of teaching effectiveness.

VI. Review of the Validity of the Findings and Conclusions

It is not possible to fully determine the validity of the report's findings and conclusions due to a marked lack of detail about statistical results and methods. For example, the MET project in general (and this report in particular) uses the Tripod survey as the primary exemplar of using student surveys to measure teaching effectiveness. However, the report contains no evidence—and includes no citations to evidence—about the psychometric properties of the scales measured by this instrument, about the reliability of subscales, about the results of any pilot work or cognitive interviews that might have established respondents' understanding of survey items or beliefs that items validly measure intended constructs, or any other empirical evidence of the instrument's validity.

These are not merely academic concerns.

- On page 15, the report discusses how the Denver Public Schools used a version of Tripod that included three items per construct. The authors indicate that Denver is assessing the consequences of this measurement decision: "The school system is assessing results to determine the reliability of the streamlined tool and the value teachers see in the feedback provided" (p. 15). Providing this background does not sufficiently alert readers to the considerable likelihood that using only three items per construct might produce measures that are too unreliable to support decisions about a teachers' effectiveness in a particular domain.

- On page 16, the authors suggest the following strategy for increasing survey reliability: "…averaging together results from different groups of students for the same teacher would reduce the effects of any variance due to the make-up of a particular class." This possible benefit is not weighed against the diagnostic information that's lost about individual classes. More broadly, this recommendation is uncritical about the potentially problematic notion that it is meaningful and valid to measure teaching effectiveness by averaging across individual classes that might differ in important, qualitative ways.

- Page 18 contains this advice:

  > Meaning comes from specificity, points of reference, and relevance. It's of little help to a teacher to be told simply, "you scored a 2.7 out of 4.0 on 'Care.'" To understand what that means requires knowing each question within the construct, one's own results for each item, and how those results compare with those of other teachers.

  Providing a valid comparison point for a teacher is not an uncomplicated issue, and more detail is needed about how this might actually be accomplished.

VII. Usefulness of the Report for Guidance of Policy and Practice

By describing specific ways that student surveys can complement other measures of teaching effectiveness and guidelines for administering student surveys, *Asking Students about Teaching* contains useful information for policy and practice. Student surveys are a useful tool in practitioners' and policymakers' toolkits, and many of the practical pieces of advice offered in this report are sensible and worth further investigation and, in many cases, worth putting into practice.

A major limitation of the report is its failure to provide sufficient supporting evidence of its claim of a relationship between student survey reports and teacher effectiveness. A broader limitation of the report is that many of its findings and conclusions are presented too uncritically and without sufficient justification. There is a school of thought that all methods are fallible and that all have unique strengths and weaknesses.[12] Developers of the MET project appear to embrace this idea with their acknowledgement that multiple measures of teaching effectiveness are needed to represent such a complex, multi-faceted phenomenon. This is a reasonable starting point. However, this report's stance is lopsided in its discussion of the potential uses of student surveys, placing too much weight on the strengths of student surveys and not enough weight on their weaknesses. A potential concern is that policymaking readers will pursue a glib implementation of some of the report's recommendations, based on an unwarranted overconfidence in student survey results.

Notes and References

1 Bill & Melinda Gates Foundation (2012). *Asking Students about Teaching: Student Perception Surveys and Their Implementation*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved November 13, 2012, from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.

2 For a review of this report, see

Rothstein J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved October 28, 2012, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

3 For a review of the report, see

Guarino, C. & Stacy, B. (2012). *Review of "Gathering Feedback for Teaching."* Boulder, CO: National Education Policy Center. Retrieved October 28, 2012 from http://nepc.colorado.edu/thinktank/review-gathering-feedback.

4 Tripod is a student survey and is one of the five candidate measures of teacher effectiveness investigated by the MET project (the others being student achievement gains, classroom observations and teacher reflections, teachers' pedagogical content knowledge, and teachers' perceptions of working conditions and instructional supports). Tripod is designed to measure "teaching, student engagement, school norms, and student demographics. To measure teaching, the survey groups items under seven constructs, called the '7 Cs': Care, Control, Challenge, Clarify, Confer, Captivate, and Consolidate" (p. 5).

5 Bill & Melinda Gates Foundation (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.* MET Project Research Paper. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved October 28, 2012, from http://www.metproject.org/reports.php.

6 Kane, T.J. & Staiger, D.O., *et al.* (2012, Jan.). *Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains.* MET Project Research Paper. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved October 28, 2012, from http://www.metproject.org/reports.php.

7 According to the first MET report (Bill & Melinda Gates Foundation, 2010), on the secondary survey, the categories were labeled "totally untrue", "mostly untrue", "somewhat", "mostly true", "totally true". On the elementary survey, the 5 choices were "no, never", "mostly not", "maybe/sometimes", "mostly yes", "yes, always".

Bill & Melinda Gates Foundation (2010) *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project.* MET Project Research Paper. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved October 28, 2012, from http://www.metproject.org/reports.php.

8 For a useful general purpose resource for questionnaire design see

Bradburn, N.M., Sudman, S. and Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: Jossey-Bass

A comprehensive guide to assessing the validity of survey instruments can be found in Chapter 8 of:

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). Survey Methodology, 2nd Edition. Hoboken, NJ: Wiley.

A useful guide for how to conceptualize and assess the validity of quantitative measures

Wilson, M. (2005). Constructing Measures: an Item Response Modeling Approach. New York, NY: Taylor & Francis Group.

9 Rothstein J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved October 28, 2012, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

10 Guarino, C. & Stacy, B. (2012). *Review of "Gathering Feedback for Teaching."* Boulder, CO: National Education Policy Center. Retrieved October 28, 2012 from http://nepc.colorado.edu/thinktank/review-gathering-feedback.

11 Having set forth this concern, I should note the possibility that the analysis might in fact have been conducted on teachers in the bottom and top quartiles of the Tripod survey distribution (as with the regression analysis discussed previously) and that the two groups of teachers were simply mislabeled in the report.

12 See for example

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

DOCUMENT REVIEWED:	**Asking Students about Teaching: Student Perception Surveys and Their Implementation**
AUTHOR:	The Bill and Melinda Gates Foundation
PUBLISHER/THINK TANK:	The Bill and Melinda Gates Foundation
DOCUMENT RELEASE DATE:	September 2012
REVIEW DATE:	November 15, 2012
REVIEWER:	Eric M. Camburn, University of Wisconsin-Madison
E-MAIL ADDRESS:	camburn@wisc.edu
PHONE NUMBER:	(608) 263-3697

SUGGESTED CITATION:

Camburn, E.M. (2012). Review of "Asking Students About Teaching: Student Perception Surveys and Their Implementation" Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-asking-students