Dimensionality and Generalizability of the Mathematical Quality of Instruction Instrument

Ben Kelcey, Dan McGinn, Heather Hill, and Charalambos Charalambous

University of Cincinnati, Harvard University, Harvard Graduate School of Education,

University of Cyprus

Abstract

The purpose of this study was to investigate three aspects of construct validity for the Mathematical Quality of Instruction classroom observation instrument: (1) the dimensionality of scores, (2) the generalizability of these scores across districts, and (3) the predictive validity of these scores in terms of student achievement.

Keywords: classroom assessment, mathematics domain, item response theory, multidimensional, validity

Dimensionality and Generalizability of the Mathematical Quality of Instruction Instrument

Classroom observations have long been viewed as a central measurement strategy for evaluating and developing teachers (Gitomer, 2009). Classroom observations, carried out for the purpose of studying practice, offer a promising way to evaluate teaching because they anchor assessments in specific and observable criteria. The value of classroom observations in identifying effective teachers and practices, however, depends heavily on the validity of the constructs measured by the observation system. Though research on the validity of these constructs varies across instruments and subjects, in no area is the evidence describing their validity potentially less complete than in mathematics.

We contribute to the literature describing the validity of classroom observations of mathematics instruction by investigating three key aspects of validity for the Mathematical Quality of Instruction (MQI) instrument. First, we examine the extent to which there is empirical support for a multidimensional structure of teaching quality. The MQI instrument describes teaching quality using indicators organized into four primary domains. Theory, prior research, and practical use of the instrument have suggested that these domains form three distinct dimensions. Our investigation assesses the relative and absolute fit of this structure and its comparative fit against other plausible structures.

Second, we investigate the extent to which the measurement of teaching quality is invariant across districts. Questions concerning the generalizability of classroom observation scores arise in both research and practice. For instance, many research questions examine relationships between observation scores and student achievement across multiple districts. Similarly, many forthcoming state policies intend to evaluate teachers from different districts using a common instrument. In each case, an important assumption underlying the validity of these assessments is that the instrument is measuring the same latent construct similarly for each district—that is, measurement is invariant across districts. However, because districts have, for example, different student populations and curricula, there may be important differences in how observation instruments function across districts.

Finally, we examine the predictive validity of teacher quality scores. A guiding principle in the theory of classroom observations is that instruments should measure teaching quality as it relates to student cognitive development. As a result, a primary benchmark for the validity of classroom observations is their efficacy in predicting student achievement gains. For this reason, we assess the predictive validity of the observation scores by correlating them with teacher value-added scores.

Methods

This study is based on data from 293 fourth- and fifth-grade math teachers, from the overarching study, and their students in five districts. Each teacher was observed and rated by two (of 39) raters using the MQI instrument (Hill et al., 2008). To investigate the dimensionality and generalizability of constructs measured by the instrument, we drew on multilevel factor analysis to compare the relative performance and fit of the different specifications.

Results

Results suggested that teaching quality was multidimensional and that the three-dimensional structure demonstrated good fit and outperformed competing structures. In terms of generalizability, we found that measurement was only partially invariant across districts. Further, each dimension significantly predicted teachers' value-added scores.

Implications

Recent initiatives have charged states with differentiating among teachers in terms of their teaching quality (U.S. Department of Education, 2011). To meet this charge, evaluators frequently employ one-dimensional descriptions of teaching quality and assume this description is invariant across districts. Our results suggest that although classroom observations offer valid descriptions of teaching as it relates to student achievement, one-dimensional assessments of teaching quality may be incomplete and the value and meaning of even multidimensional descriptions may vary across districts.

References

Gitomer, D. (2009). Measurement Issues and Assessment for Teaching Quality. London: Sage Publications.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. Cognition and Instruction, 26, 430–511.

U.S. Department of Education (2009). Race to the Top Fund – ExecutiveSummary: Notice of Proposed Priorities, Requirements, Definitions, and Selection Criteria. Washington, DC: U.S. Department of Education.