Using Item Response Theory to Learn about Observational Instruments

Dan McGinn, Ben Kelcey, and Heather Hill

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

Abstract

As many states are slated to soon use scores derived from classroom observation instruments in high-stakes decisions, developers must cultivate methods for improving the functioning of these instruments. We show how multidimensional, multilevel item response theory models can yield information critical for improving the performance of observational instruments.

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

Using Item Response Theory to Learn about Observational Instruments

In the policy landscape of Race to the Top and No Child Left Behind waivers, many states are in the process of implementing high-stakes teacher evaluation systems. A teacher's evaluation in these systems depends in part on scores from classroom observation instruments, and so teachers, administrators, states, and unions all have a vested interest in ensuring these scores are as valid and reliable as possible. However, few tools are available for evaluating the performance of observational instruments, leaving developers with few avenues to improve their instrument. One tool that has been used to examine observational instruments is generalizability theory (Hill, Charalambous & Kraft, 2012). While yielding more nuanced estimates of reliability than Cronbach's alpha, generalizability theory does not yield estimates of the difficulty of items, nor describes their ability to discriminate among teachers. We illustrate how multidimensional, multilevel item response theory (IRT) models can be used to gain valuable information about the functioning of observational instruments, with data collected using the Mathematical Quality of Instruction (MQI) instrument (Hill et al., 2008).

We investigate the dimensional structure of the MQI, and find support for three dimensions. We then use a three-dimensional, two-level, graded response model, with scores for each 7.5 minute segments nested within teacher. We use the output from these models to evaluate items on the MQI. We first investigate item discrimination parameters to ensure that items are related to the underlying construct as theorized. We then see that many difficulty parameters associated with item score points are extremely high (from 4 to 8 or 9 standard deviations above average), yielding little information for average or below average teachers.

Guided by these findings in Year 1 data, we inserted additional, easier to endorse score points in selected items. We scored both the original and expanded versions in Year 2, and discuss the results.

We conclude that some practices valued by math educational researchers are rarely enacted in the broad population of teachers, which leads to suboptimal measurement properties. Multidimensional, multilevel IRT models have been shown to be appropriate choices for instruments with seldom-endorsed items (Raudenbush, Johnson & Sampson, 2003), and yield valuable information with which developers can improve their instrument. Such techniques will be of increasing importance as practices described by the Common Core State Standards, the guide for many observational instruments now in development, are likely to be infrequently enacted.

References

Raudenbush, S., Johnson, C., Sampson, R. (2003). A multivariate, multilevel Rasch model with application to self-reported criminal behavior. *Sociological Methodology 33*(1), 169-211.

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