Teacher evaluation and the Standards of Effective Instruction rating instrument: Psychometric considerations

Christopher Moore

MINNEAPOLIS PUBLIC SCHOOLS
Urban Education. Global Citizens.

Introduction

Minneapolis Public Schools (MPS) and the Minneapolis Federation of Teachers (MFT) are collaboratively developing a teacher development and evaluation process. The process will comply with Minnesota Statutes 2011, 122A.40, subd. 8: "To improve student learning and success, a school board and an exclusive representative of the teachers ... may develop a teacher evaluation [process] through joint agreement... The process must include [1] having trained observers... [2] value-added assessment model ... as a basis for 35 percent of teacher evaluation results ... [3] longitudinal data on student engagement and connection."

During the 2012-2013 school year, each teacher will receive feedback from two observers over five occasions. Observers have been trained and certified to use the Standards of Effective Instruction (SOEI) rating instrument comprising 30 key items. All of the items will be scored during at least one occasion; a subset of items will be scored on most occasions. Item are scored as 1 (requires attention), 2 (developing), 3 (proficient), or 4 (exemplary).

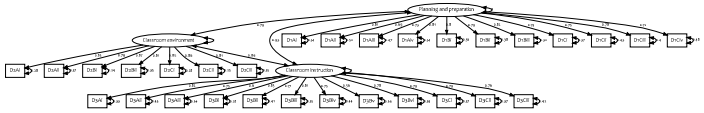
Research questions

Using scores from the 2011-2012 pilot year, this study addresses two research questions:

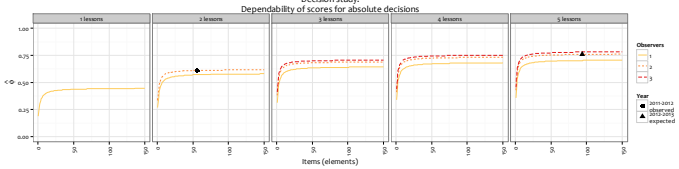
- Does the SOEI measure an essentially unidimensional "effective instruction" construct as theorized? If not, then observers will need to be trained to score several items from each latent factor on every occasion in order to ensure content validity and adequately reliable scores.
- To what degree were SOEI scores reliable? Reliable/precise scores are necessary to infer a teachers' instructional effectiveness level and track improvements over time.

Methods and results

I conducted a confirmatory factor analysis to assess the degree to which the SOEI instrument measures a single latent factor. A polychoric correlation matrix based on 704 observed lessons was analyzed. For teachers observed more than once, only scores from the first occasion were included. I specified two models: a single "effective instruction" factor solution that holds practical benefits and a less parsimonious three-factor/domain solution. In the latter specification, the SOEI's three domains (planning and preparation, classroom environment, and classroom instruction) represent separate but correlated latent factors. The three-factor solution (see below) exhibited better fit as measured by the adjusted goodness of fit index (0.71 vs. 0.59), Bayesian information criterion (3917.3 vs. 5076.3), and standardized root mean square residual (SMSR; 0.05 vs. 0.06). The single-factor solution cannot be ruled out because the three factors were highly correlated and both models suffice according the SMSR criterion of 0.10 (Kline, 2005). **Taken together, these results suggest that total scores based on all items can be reported for practical benefit but domain scores should be reported, too.**



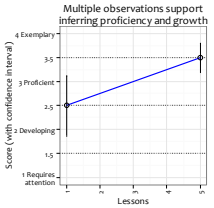
I conducted a generalizability study in order to estimate the degree to which SOEI scores are reliable. Generalizability theory is appropriate when observed scores are composed of error due to measurement conditions (e.g., raters) in addition to random error and a person's true score of interest (Shavelson & Webb, 1991). **As shown in the decision study plot below, scores based on 56 items rated by two observers separately over two occasions/lessons yielded a reliability coefficient of 0.61. In the 2012-2013 school year, teachers will receive scores based on about 94 items rated by two observers separately over five occasions/lessons, which is expected to reach a reliability level of 0.76.** For perspective, reliabilities of student scores from Title I assessments regularly exceed 0.90, and the Measures of Effective Teaching project of the Bill and Melinda Gates Foundation (Kane & Staiger, 2012) recommends a minimum level of 0.67 for teacher observation scores.



Discussion

MPS' teacher evaluation process needs to be fair and support teachers' development. It also needs to allow inferences about where teachers are in terms of instructional effectiveness so improvements can be tracked over time. Support and inference are not mutually exclusive. They are bound together by score reliability. A single score from one rater who observed one lesson offers limited usefulness and reliability. Only after multiple observations can a teacher begin to see how their efforts, informed by scores from earlier observations, are paying off. Additionally, as more observations are conducted, score reliability and precision increase to a point where inferences can be made.

The importance of reliability is illustrated for a hypothetical teacher in the plot below. In the first round of observations the teacher scored a 2.5 with a large confidence interval, but by the fifth round we can infer that the teacher had reached proficiency and improved significantly because the confidence intervals do not overlap. **Given the importance of reliability, it should be a key metric for evaluating teacher evaluation.**



References

Kane,T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. Guilford press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.