Examining the Validity of Observation Scores of Special Education Teachers in High-Stakes Teacher Evaluation Systems

Despite all 50 states including observation systems in teacher evaluation systems, no research studies have examined the validity of general observation systems for use with special educators, a group that represents approximately 14.5% of the teaching workforce. Therefore, the purpose of this study is to examine whether the Framework for Teaching (FFT), an observation scheme that has been widely adopted across the country, reliably and validly captures special education teachers' instruction.

Authors:

Presenting Author: Nathan Jones, Boston University (ndjones@bu.edu)

Courtney Bell, ETS (cbell@ets.org)

Mary Brownell, University of Florida (mbrownell@coe.ufl.edu)

Yi Qi, ETS, (yqi@ets.org)

Background and Purpose

All 50 states now require classroom observation systems in teachers' evaluation systems, and there is evidence that schools are using the information derived from observations – more than from student learning outcomes – to foster teacher development and to guide human capital decisions (Goldring et al., 2015). However, despite the widespread adoption of observation systems in evaluation, no research studies have examined the validity of general observation systems for use with special educators, a group that represents approximately 14.5% of the teaching workforce. Further, researchers have raised concerns about whether such systems appropriately account for the kinds of instructional practices most commonly used by special educators; specifically, how do observation tools designed to measure the quality of student-centered instructional practices function for a population of teachers who have been trained to use explicit, teacher-directed approaches to instruction (Jones & Brownell, 2015)?

If observational data are to fulfill their promise – to distinguish between more- and less-effective teachers and to help those who are struggling to improve – then it will be critical to have empirical evidence of whether these data provide reliable, valid data on the instruction of important sub-groups of educators. Therefore, the purpose of this study is to examine whether the Framework for Teaching (FFT), an observation scheme that has been widely adopted across the country, reliably and validly captures special education teachers' instruction.

To validate the FFT for use with special educators, we adopt Kane's validity argument approach, in which we appraise the plausibility of the argument that judgments of special educator teaching quality can be made on the basis of FFT scores. We appraise the plausibility of the argument that judgments of special educator teaching quality can be made on the basis of FFT scores through an empirical evaluation of four related sets of inferences: 1) scoring, 2) generalization, 3) extrapolation, and 4) interpretation. Drawing on a sample of approximately 320 videotaped lessons from 80 special education teachers in Rhode Island and Idaho, we examine validity evidence on the first three of these inferences.

Setting, Participants, and Data Collection

We collected classroom observation data from 80 special education teachers in Rhode Island (N=51) and Idaho (N=29) in the 2016-2017 school year. Each participating teacher was observed four times providing instruction in reading, math, or both if relevant. We included elementary and middle school special education teachers in grades 3-8, targeting instruction provided in co-teaching and teaching in resource rooms to high-incidence student populations. Participants also completed surveys and were interviewed about their backgrounds, teaching context, and evaluation experiences. Administrative data on teachers (years of experience, certification, evaluation history) were collected through the state departments of education.

Once videos were collected, each video was scored using two instruments – the FFT and an observation instrument more closely aligned with special education teaching practices, the Classroom Observation Student-Teacher Interactions (COSTI) Instruction (QCI) (Doabler et al., 2014; 2015). To ensure comparability of score quality, parallel training regimes were used with each instrument; i.e., raters received a systematic four-day training followed by a certification

exercise to ensure that raters were ready to score in practice. Raters completed their scoring over a five-week period and completed weekly calibration exercises to prevent rater drift. To examine consistency of scores across raters, all videos were double scored on each instrument.

Research Design and Analysis

Validation activities are conducted for each set of inference's in Kane's validity framework. For the *scoring inference*, we examine the relationships among dimensions on FFT as well as the internal reliability of scores. We assess model fit (i.e., whether the expected domain structure is supported by the data) by conducting confirmatory factor analysis (CFA). However, given the relatively small sample size of 80 teachers, the adequacy of the sample for CFA will depend on the correlation of our two factors. We also examine the agreement of raters' scores to one another and assess rater bias by examining whether any raters systematically score differently than others. For the *generalization* inference, we conduct G studies to examine variation in lessons by teacher, rater, lesson, day, content, type of special education service delivery model (i.e., resource instruction, co-teaching instruction), and residual error. Variance components are estimated using standard random effect analysis of variance methods (Searle, Casella, & McCullough, 1992) Finally, for the *extrapolation* inference, we examine whether FFT scores are correlated with data from other measures of high quality special education teaching, including teachers' value-added scores (if teachers have adequate numbers of students to calculate scores) and scores drawn from the COSTI (to correct for measurement error, disattenuated correlations are used). We calculate Spearman's rank correlation coefficients and Kendall's τ to determine whether the measures rank order teachers similarly.

Findings

Videos were scored in July and August, 2017. Analyses will be conducted prior to the March SREE conference. Preliminary evidence suggests that a large proportion of the variation in scores is attributable to raters and to instructional setting (co-taught classrooms vs. self-contained/resource settings). Also, overall, a large proportion of the teachers in the sample appear to score low on the FFT and the COSTI, raising questions about the quality of special education teaching observed in practice. This finding aligns with data from other recent studies of special education teaching.

Conclusions

There are complex conceptual and logistic challenges in adopting a common evaluation tool across all subgroups of teachers in a state. While our results examine the case of special educators, it is easy to imagine other populations of teachers for whom similar challenges might emerge. In the absence of empirical data on the FFT's validity with various subgroups of teachers, it will be difficult to know whether the instrument can help improve teachers' instruction. Therefore, although there are limitations in our sample (namely that we draw on teacher volunteers from multiple districts in two states), few previous studies have examined videotaped classroom data for similarly large samples of special education teachers, and no studies have used the FFT with these teachers. Therefore, empirical findings like these will be

critical in determining whether existing observation systems can be used reliably and validly in the evaluation of special educators.

References

Doabler, C. T., Nelson, N. J., Kosty, D. B., Fien, H., Baker, S. K., Smolkowski, K., & Clarke, B. (2014). Examining teachers' use of evidence-based practices during core mathematics instruction. *Assessment for Effective Intervention*, *39*(2), 99-111.

Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *The Elementary School Journal*, *115*(3), 303-333.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make Room Value Added Principals' Human Capital Decisions and the Emergence of Teacher Observation Data. *Educational Researcher*, *44*(2), 96-104.

Jones, N. D., & Brownell, M. T. (2015). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39*(2), 112-124.