Does Special Educator Effectiveness Vary Depending on the Observation Instrument Used?

Evelyn S. Johnson, Angela R. Crawford, Yuzhu Zheng, and Laura A. Moylan

Boise State University

Author Note

Evelyn S. Johnson, Department of Early and Special Education, Boise State University; Angela Crawford, Project RESET; Boise State University; Yuzhu Zheng, Project RESET, Boise State University, Laura A. Moylan, Project RESET, Boise State University.

This research was supported the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors.

Correspondence regarding this manuscript should be addressed to: Dr. Evelyn Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

Citation: Johnson, E. S., Crawford, A. R., Zheng, Y. & Moylan, L. A. (2020, in press). Does Special Educator Effectiveness Vary Depending on the Observation Instrument used? *Educational Measurement: Issues & Practice*.

Abstract

In this study, we compared the results of 27 special education teachers' evaluations using two different observation instruments, the Framework for Teaching (FFT), and the Explicit Instruction observation protocol of the Recognizing Effective Special Education Teachers (RESET) observation system. Results indicate differences in the rank-ordering of teachers depending on which instrument was used. Overall scores on RESET were higher on average than those on FFT. Item level analyses showed that across 125 correlations, 73 were significant, low-moderate, and 52 were non-significant. Implications for research and practice are discussed.

Does Special Educator Effectiveness Vary Depending on the Observation Instrument Used?

Observations of teacher practice are designed to provide direct measures of teaching effectiveness through the the use of protocols that capture the salient elements of instruction. *Effective* observation protocols provide information that can be used to evaluate, identify and improve specific teaching practices (Gitomer et al., 2014). An ongoing tension in the field of teacher observation is whether there is a need for subject specific observation tools, or whether instruments that reflect general instructional practices across grade levels and content areas can provide teachers sufficient information to change their practice. As the focus of teacher observation has shifted largely from accountability towards instructional improvement, there has been a growing call for the development and use of subject specific observation instruments that provide concrete guidance and detailed, actionable feedback to teachers on the instructional practices identified as most effective for the specific area they teach (Blazar et al., 2017; Hill & Grossman, 2013).

Within the field of special education, it has been argued that special education specific observation tools are needed not only to give teachers more detailed feedback but also because general instruments are not aligned with the evidence-based practices (EBP) that support higher achievement for students with disabilities (SWD; Johnson & Semmelroth, 2014; Jones & Brownell, 2014). For example, in a detailed analysis of Danielson's Framework for Teaching (FFT; Danielson, 2013), Jones and Brownell (2014) report that "…the core responsibilities of the special educator may not be represented in the Framework." (p. 119). In particular, they note that explicit instruction, a high leverage practice (HLP; McLeskey et al., 2017) for students with

disabilities is absent from FFT, and that the scoring criteria to achieve a "distinguished" rating on FFT is counter to most of the characteristics of explicit instruction (Jones & Brownell, 2014).

In a study of 84 special education teachers evaluated using the instructional domain of FFT, the mean scores across the five components ranged from 1.78 to 2.3 (Jones, 2019). FFT is rated on a four-point scale, where a 1 is unsatisfactory, a 2 is basic, a 3 is proficient, and a 4 is distinguished. Initial conclusions from this study include that the FFT "does not seem to pick up on common practices used with SWDs" (Jones, 2019), and that a special education teacher observation system that can successfully affect change in the classroom must be based on the systematic measurement of the implementation of EBPs to support the needs of SWD (Johnson & Semmelroth, 2014). However, without scores from an instrument aligned with special education practices, it is impossible to discern whether the low performance in their sample reflects poor teaching ability, rater severity, or a misalignment of the observation protocol with the desired instructional practice.

Recent research of teacher evaluation for subject areas other than special education supports the idea that subject specific observation tools capture important constructs related to teacher effectiveness and student outcomes. For example, ongoing studies of the Mathematical Quality of Instruction (MQI) instrument have found that "there is a powerful relationship between what a teacher knows, how she knows it, and what she can do in the context of instruction" (Hill et al., 2008, p. 496). Analyses of the Protocol for Language Arts Teacher Observation (PLATO; Grossman et al., 2009) have found strong relationships between teacher performance and student outcomes on various assessments (Grossman et al., 2014). Emerging evidence in special education indicates that after controlling for students' baseline scores, teacher performance on an explicit instruction observation protocol accounted for unique variance

(4.4%) in student growth on standardized, curriculum-based measures in reading and math (Johnson et al., 2020).

Despite these arguments in favor of subject specific measures, there are considerations of implementation that must be considered. First, the development of subject specific measures is not an insignificant undertaking. Although there are current tools and related research efforts to develop observation tools in math (MQI, Hill et al., 2008), language arts (PLATO; Grossman et al., 2009), and special education (RESET; Johnson et al., 2018a), there remain a number of subject areas for which there are no existing observation instruments. Second, given the extensive training required to achieve reliability among raters, it is more cost-effective to train multiple raters to assess instruction using one instrument as opposed to multiple instruments (Jones & Brownell, 2014). There is some evidence to suggest that raters with expertise in special education are more accurate in evaluating special education teachers than those without. Lawson and Cruz (2018) reported that when building level administrators with limited special education experience rated a set of special education teaching videos, the variance attributable to raters in a generalizability study was 20.9%. This percentage dropped to .2% when ratings were conducted by special education teachers with more than five years of experience. However, finding content experts who can accurately and consistently interpret the elements included within an observation protocol will be a challenge, even in larger school districts (Hill & Grossman, 2013). Third, in surveys and reports on school leader preferences for teacher evaluation systems, many administrators report a preference for a common system, believing it leads to a more fair and equitable evaluation process and encourages a uniform vision of effective teaching (Holdheide et al., 2012; Jones & Brownell, 2014).

An important first step in deciding whether subject specific or general instruments should be used is to examine whether they lead to different conclusions about a teacher's performance. To date, there is limited research examining the extent to which evaluations of teacher practice differ based on scores provided by general or subject specific observation tools. Early studies using data collected through the Measures of Effective Teaching (MET) project found that items tended to cluster on a main factor or factors within an instrument, suggesting that different instruments capture different dimensions of teaching (e.g. Kane & Staiger, 2012; McClellan et al., 2013). However, these studies did not account for facets of the observation protocol such as raters, the number of points on the scoring scale, or the order in which various dimensions are presented, all of which may lead scores to cluster more strongly within instruments than across them (Blazar et al., 2017; Crocker & Algina, 2008). In studies that have attempted to account for observation protocol features, items from different instruments have been found to cluster onto the same factors. For example, Lockwood et al (2015) used Bayesian factor analysis to control for dimension order and found two distinct teaching constructs in their examination of general (CLASS; Pianta et al., 2008 and FFT; Danielson, 2013) and specific (MQI; Hill et al., 2008 or PLATO; Grossman et al., 2009) instruments: one for teachers' instructional practice and the other for classroom management, with items from different instruments clustering onto the same factor. Blazar et al (2017) examined whether general (CLASS) and content specific (MQI) instruments tapped different constructs using bi-factor confirmatory factor analysis to control for construct irrelevant variance and discovered a small degree of overlap across the measures, with strong evidence for factors that are distinct to each instrument.

These studies provide initial evidence that additional information about teachers' instructional practice is gained through the use of subject specific observation instruments. In the

current study, our goal is to add to this limited research by examining the extent to which special education teacher performance varies depending on whether a general observation instrument, Danielson's Framework for Teaching (FFT; Danielson, 2013) or special education specific, RESET (Johnson et al., 2018b) is used. The specific research questions addressed include:

1. Do ratings of special education teachers' practice differ systematically when evaluated using FFT as compared to RESET?
2. What is the relationship between items on the FFT and items on RESET?

Methods

Participants

Special education teachers. Twenty-seven special education teachers from three states (Idaho, Florida, Wisconsin) participated in this study. Data collection took place during the 2015–2016 and 2016–2017 school years. All participants provided video-recorded lessons that reflected their use of explicit instruction to provide reading or math intervention within a Tier 3 setting. All participants were female, teaching from first to eighth grade levels in a resource room context. Two of the 27 teachers were Asian, and the remaining 25 teachers were White. Their number of years of experience ranged from 1 to 27 years ($M = 8.4$, $SD = 3.7$). All participants held a Bachelor's degree in special education, and 16 teachers also had a Master's degree in either Special Education or Literacy Education. Participating teachers worked across a variety of school settings. Table 1 provides the demographics of the schools in which data were collected.

RESET raters. One male and fourteen female raters were recruited from seven states to score videos provided by the 27 teachers. Twelve raters were white, two Asian, and one Pacific Islander. Criteria for raters included having five or more years of experience working with

SWD. All raters were special education professionals with between 5-20 years of working experience. Two raters had a bachelor's degree in Special Education, eleven had a master's degree, and two had doctoral degrees. At the time of the study, eight raters worked as classroom teachers, three were mentor teachers or instructional coaches, two were special education graduate students, one was a specialist at a state Department of Education, and one was a school psychologist and RTI coordinator within her district.

FFT raters. Twelve raters were recruited using an electronic recruitment letter which was sent to a mailing list comprised of raters who had participated in a similar scoring project at Education Testing Services (ETS) using Danielson's Framework for Teaching (FFT). Raters were selected based on their past experience as scorers and relevant professional experience. All raters had some teaching experience. Four taught at the post-secondary level, four at the secondary level, and four at the elementary level. Four had experience teaching special education. All had at least a master's degree, and three had doctoral degrees. Four had advanced degrees in English language arts, five in educational leadership, three in elementary education, one as a media specialist, and two in special education. All raters had scoring experience, and all had prior experience scoring using FFT.

Measures

RESET Explicit Instruction observation protocol. The RESET Explicit Instruction (EI) observation protocol (Johnson et al., 2018a) consists of 25 items, with specific descriptors for each item across three levels of performance: a) 3 or implemented, b) 2 or partially implemented, and c) 1 or not implemented. The scoring scale was originally designed to include four points to align with FFT, however, performance level descriptors to meaningfully discriminate "proficient implementation" from "distinguished" have not yet been developed

(Johnson et al., 2018a). Studies examining the EI observation protocol's psychometric qualities indicate that it can serve as a reliable and consistent observation instrument of teachers' implementation of explicit instruction, with a g-study reliability coefficient of .74 (Crawford et al., 2018), and many-facet Rasch measurement (MFRM) analyses indicating strong fit statistics and reliability and separation statistics for all facets of the protocol (Johnson et al., 2018a; Johnson et al., 2018b; Johnson et al., 2019).

FFT observation rubric. The five components of Domain 3: Instruction, from the Danielson Framework for Teaching (FFT; Danielson, 2013) were used to evaluate teachers' instruction for this study. The five components include: a) Communicating with students, b) Using questioning and discussion techniques, c) Engaging students in learning, d) Using assessment in instruction, and e) Demonstrating flexibility and responsiveness. The FFT is scored on a one to four-point scale, where 1 = unsatisfactory, 2 = basic, 3 = proficient, and 4 = distinguished. Several studies have used the FFT as an indication of teacher effectiveness, but as explained by Liu et al (2019), many aspects of the way in which the FFT is used across studies vary substantially or are not reported, which makes it difficult to report reliability and validity coefficients that are meaningful across contexts.

Procedures

Video collection. All participating special education teachers were asked to video record weekly lessons with a consistent group of students using the Swivl® video capture and upload system. To decide on appropriate lessons for recording, research project staff contacted each teacher to discuss the lessons they were planning to record. Based on the information provided, the teachers then targeted a specific instructional group to record. Research staff viewed the first lesson submitted to ensure that it reflected explicit instruction. Teachers were sent a short video

and set of instructions that demonstrated how to use the Swivl® system and were provided with project staff contact information for technical support. Each teacher contributed a total of 20 video recorded lessons ranging from 20 to 50 minutes in length over the school year.

For this study, three videos from each teacher were selected from the videos they submitted, resulting in an initial lesson total of 81 videos for this study. Two video files that were scored with RESET were corrupted in the transfer to the FFT rater team and were therefore dropped from the analysis, resulting in a total of 79 lessons. Two teachers had two lessons scored, 25 teachers had three lessons scored. Observation studies have demonstrated that between two and four lessons are needed for reliable observations of a teacher's instruction (Hill, Charalambous, & Kraft, 2012; Johnson & Semmelroth, 2015), suggesting that the inclusion of two to three videos per teacher could be expected to produce reliable ratings of teachers' performance. Videos were first organized into three categories based on time of year. Videos from the first nine weeks were considered the beginning of the year, videos from the second nine weeks were considered the middle of the year, and videos from the last nine weeks were considered the end of the year. One video from each time category for each teacher was randomly reviewed to ensure adequate audio and video quality and selected to be evaluated. These video selection criteria (e.g., three observations, across a school year) were adopted because they are consistent with how an administrator would typically use the observation system. Each video was assigned an identification number and listed in random order for each rater to control for order effects.

RESET rater training. Over a four-day training period, raters were first provided with an overview of the RESET project goals and a description of how the EI rubric was developed. Research project staff then explained each item of the EI rubric and clarified any questions the

raters had about the items. Raters were also provided with a training manual that included detailed descriptions of each item, along with examples for each item across each level of performance. Then, raters watched and scored a video that had been scored by project staff. The scores were reviewed and discussed to include the rationale for the score that each item received. Raters then watched and scored three videos independently, and scores were reconciled with the master coded rubric for each video. Any disagreements in scores were reviewed and discussed. To determine rater agreement, Kendall's coefficient of concordance, W, was used to allow for ordinal data with multiple raters. For the first training video, $W = .191$ $p < .001$. For the fourth video, $W = .303$, $p < .001$. During the training, minimum rater performance standards were not established. Some studies have shown that raters still account for large portions of variance, and issues, such as drift, persist even with establishing minimum performance standards (Cash et al., 2012; Jones, 2019; Kane & Staiger, 2012.). Instead, research project staff communicated regularly with raters to support their understanding of the items. MFRM allows us to investigate the internal consistency of the raters and adjusts parameters of items, teachers, and lessons for discrepancies in severity of raters.

After training, raters were then assigned a randomly ordered list of videos and asked to evaluate the videos following the assigned order, to score each item, to provide time stamped evidence that they used as a basis for the score, and to provide a brief explanation of the rationale for their score. Raters were reminded to consult the training manual as they completed their observations and were given a timeframe of four weeks to complete their ratings. Completed evaluations were submitted using an electronic version of the rubric developed in the Qualtrics ® survey system. Although no calibration process was involved during the scoring, RESET research staff checked raters' progress and data entry every week. Through emails and phone

calls, the research staff helped raters who had some difficulty in understanding and applying the rubrics to rate the videos.

To maintain a feasible video observation load for each rater, a rating scheme was developed to link scores across raters and videos without requiring each rater to score each video (Eckes, 2011). Two teachers were randomly selected to have their first and last video scored by every rater. Remaining videos were randomly assigned so that each video was scored by four raters. This created a design in which 13 raters scored 28 videos each, one rater scored 32 videos, and one rater scored 8 videos.

FFT rater training. The 79 videos used for this study were part of a larger research effort on FFT conducted by ETS. Twelve raters attended the in-person training on the ETS campus in Princeton, NJ. The training spanned five full days and focused on the application of the framework in general, with particular focus on those aspects that historically prove challenging for raters. After the framework was deeply reviewed, raters practiced scoring as a group, then individually, with trainer guidance and feedback. Raters needed to pass a certification test before being allowed to participate in operational scoring. To pass a certification test, raters were required to score two full-length lessons using the FFT and achieve the established standard of accuracy. The first certification test took place on the last day of the in-person training. Certification results were calculated, and raters were notified via email. Seven raters passed the first certification test. The five raters who didn't pass the first certification test were required to attend a remediation phone call two or three days after the first certification test. During the remediation calls, trainers explained the rationale and justification of master scores, answered raters' questions, and clarified important concepts of the FFT instrument. Subsequently, these five raters took a second certification test, which also involved scoring two

full-length lessons. All raters passed the second certification test and were allowed to proceed to operational scoring.

Videos were rated over a nine-week period. Scores were submitted once a week and checked for completeness, proper segmentation, and errors in data entry. Raters scored the calibration video each week until all scoring was complete. After the calibration scores were submitted, project staff reviewed the scores and shared them with the rater trainers to inform the weekly call. Raters did not need to maintain a specified level of agreement with master scores in order to continue in the study.

Data Analyses

Data were analyzed in multiple ways to address the research questions. The evaluation of teacher practice with any observation protocol is complex, with multiple facets including raters, items, lessons, and time of year shown to impact performance (Casabianca et al., 2015; Hill et al, 2012; McCaffrey et al., 2015). This complexity makes it extremely challenging to examine how different instruments might lead to different conclusions about a teacher's performance. Even when methods such as certification, calibration and ongoing validation are used to limit rater error, exact agreement levels ranging from 47 – 52% across a number of instruments (including FFT and RESET) have been reported (Casabianca et al., 2015; Cash et al., 2012; Johnson et al., 2019; Jones, 2019).

Research on rater behavior suggests that achieving perfect agreement is an elusive goal and that acknowledging that raters will differ in their severity but can be trained to be consistent in their own scoring may be a more attainable reality (Eckes, 2011). Many-faceted Rasch measurement (MFRM; Linacre, 2014), is an approach to data analysis that allows for the investigation of multiple facets (e.g. teachers, lessons, items, raters) to understand how they

function within the measurement process, and to examine their interactions. An advantage of using MFRM to analyze teacher observation data is that it computes a "fair average score" that controls for the various facets that impact a teacher's performance. Therefore, to mitigate the effects of variability due to rater severity, we first used MFRM analyses to compute fair average scores (Linacre, 2017) for evaluation scores provided on both RESET and FFT.

MFRM analysis. The model used for the MFRM analysis in this study is given by:

$$\ln P_{nijok} P_{nijo(k-1)} = B_n - D_i - C_j - T_o - F_k$$

where P_{nijok} is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*. $P_{nijo(k-1)}$ is the probability of teacher *n*, when rated on item *i* by judge *j* in occasion *o*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_j is the severity of judge *j*, T_o is the stringency of occasion *o*, and F_k is the difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011). The MFRM analysis was conducted using the computer program FACETS version 3.80 (Linacre, 2017). For this study, we used the fair average score computed both at the teacher level and at the lesson level for both FFT and RESET scores.

Descriptive, correlational and statistical analyses. Once the fair average scores were computed, we used them to rank order teachers' performance on each instrument, at both the teacher and the lesson levels. We computed a Spearman's rank order correlation and a Pearson's correlation coefficient at both the teacher and lesson levels to examine the relationship of these variables. Next, we examined the relationship of items on RESET with the five components of FFT's Domain 3. Finally, we ran t-tests at both the teacher and lesson levels to determine whether scores based on RESET systematically differed from those obtained with FFT.

Results

Teacher Level Results

First, we examined scores at the teacher level, with results presented in Table 2. Teachers were all assigned a numerical identifier, and the first column of Table 2 presents the rank order of teachers based on their RESET fair average score, reported in Column 2. Their corresponding FFT fair average score is in Column 3, and the resulting rank order based on FFT is presented in Column 4. Twenty-three of 27 (85%) teachers had higher fair average scores with the RESET instrument, mean difference = .29, *SD* = .18. Three teachers (Teachers 54, 16, 57) had higher fair average scores with the FFT, and one teacher's fair average score was the same on both instruments (Teacher 9). The Pearson correlation coefficient between the two scores was 0.63 (p <. 01). To better illustrate the difference in rank order based on instrument, we divided the teachers into quartiles and highlighted each with a different color based on the RESET rank order. The third column retains the RESET quartile highlighting to show where the teachers within each quartile are ranked with the FFT instrument. The Spearman rho of -0.05 indicates there is no correlation between the rank orders. Two teachers (Teacher 7, Teacher 5) had the same rank with both instruments. Eleven of 27 teachers fall within the same quartile on both instruments.

To determine whether there were statistically significant differences between teachers' performances on the two instruments, we conducted a paired samples t-test using the teachers' fair average score, $df = 26$, $t = 5.567$, $p < .001$. In addition to the statistical difference, there are important practical differences. As can be seen in Table 2, on RESET, three teachers or 11% of the sample (Teachers 10, 8, 52) have a score below 2.0, which places their performance in the "not implemented" category. On FFT, an additional six teachers for a total of 9, or 33% of the

sample had scores that place their performance in the "unsatisfactory" category, including Teacher 26, who was in the top quartile when scored using RESET.

Item Level Results

A total of 1975 scores (25 items x 79 lessons) were assigned to the instructional lessons using the RESET rubric, and a total of 2247 scores (five components x 79 lessons x varying numbers of segments) were assigned using FFT. Table 3 presents the score distribution for each instrument. Whereas fewer than 10% of items received a score of "not implemented" using RESET, 17% of the scores assigned with FFT fell were "unsatisfactory." Additionally, 40% of items received a score of "proficient implementation" with RESET, whereas only 23% received a "proficient rating" with FFT.

Table 4 presents the correlations between items on each instrument. Several items on RESET and FFT are statistically significantly correlated. To better understand these correlations, Table 5 presents the item descriptions at the "proficient" level for both RESET and FFT Domain 3. Component 3a on FFT, "Communicating with students" was significantly correlated with nearly every item on RESET with the exception of items 17 "allowing think time" and 23 "providing feedback". Component 3b on FFT, "Using questioning and discussion techniques" had statistically significant correlations with 12 RESET items. Component 3c on FFT, "Engaging students in learning", had statistically significant correlations with 16 RESET items, and Component 3d, "Using assessment in instruction", with 19 RESET items. Component 3e, "Demonstrating flexibility and responsiveness" was only significantly related to 3 RESET items. There are only 4 items across instruments that have correlation coefficients greater than .40, with the largest correlation of .48 between FFT Component 3c and RESET item 22, "Teacher consistently checks for understanding".

Discussion

Results from this study suggest that differences in judgments made about special education teachers' instructional performance emerge when they are evaluated using a content-specific (RESET) as opposed to a general observation (FFT) instrument. General observation systems like FFT are being used in a number of states to make important decisions about teachers' practice, yet limited research examining the appropriateness of the scoring decisions, especially as they pertain to special education teachers, is available (Herlihy et al., 2014). After controlling for differences in rater severity through MFRM analyses, our findings indicate that special education teachers' instructional practice is consistently evaluated at higher levels of proficiency when an observation protocol aligned with explicit instruction, an HLP in special education is used.

Of particular concern are the differences in teacher rank order based on the instrument used. Teacher 26 presents an extreme example, in which on RESET, the evaluation suggests that this teacher is implementing explicit instruction, a HLP for SWD, with a high level of proficiency. On FFT however, Teacher 26 has a score that is just below "basic". If observers use information from the observation protocol to provide feedback to teachers, then this teacher would receive very different feedback about their level of teaching performance and what to do to improve it. Overall, there were six teachers whose performance on RESET was at a "basic" or above level, but who had FFT scores that fell below basic.

At the item level, the number of statistically significant correlations was 73 of 125 or 58%, much greater than the 5% we would expect to see by chance alone. This suggests that the instruments may capture some similar aspects of instruction. The number of statistically non-significant correlations was 52, which also suggests that there are aspects of instruction that are

different in these instruments. In this analysis, correlations at or below .20 were non-significant. In particular, FFT Component 3e, "Demonstrating flexibility and responsiveness" was correlated at statistically significant levels with only three RESET items, and was negatively correlated (albeit at very low levels) with seven RESET items. There are three RESET items (items 10, 17 and 25) that do not correlate with any component of FFT at a level greater than .30, and 13 items that do not correlate with four of the five FFT components at a level greater than .30.

Although our small sample size prevented us from conducting factor analyses to examine whether items across instruments loaded on different constructs as others have done (e.g. Blazar et al., 2017; Lockwood et al., 2015), the low correlations across items suggest that while there may be a small degree of overlap, there are also distinct aspects of instruction captured by the two instruments. Our findings lend support to the criticisms and concerns raised by Jones (2019) and Jones and Brownell (2014) that not only are important elements of special education instruction not reflected in FFT, but that there are also elements of instruction as depicted in FFT that run counter to best practice as depicted in the RESET explicit instruction rubric. If a primary purpose of teacher observation is instructional improvement, then it is critical for teachers to be evaluated and given feedback in ways that are specifically aligned with the instructional practices that have been demonstrated to be the most effective for the student population they serve. Future research should investigate whether teacher performance, as measured by RESET, accounts for more variance in student growth than teacher performance, as measured by FFT.

Limitations

There are a number of limitations in this study that call for caution in generalizing results. First, although recent reports estimate the special education workforce in general is limited in its diversity with 87% of teachers female and 81% White (Deloitte, nd), the sample of special

education teachers was small and extremely limited in its diversity. Second, although MFRM analyses help to account for construct-irrelevant variance, observation systems are comprised of rating specifications, rating processes and sampling and scoring specifications (Liu et al., 2019) that all impact the resulting assigned scores. Given the resource-intense nature of teacher observation research, especially the time and expense needed to train raters, accounting for all of these factors across any set of observation instruments remains extremely challenging. Finally, whereas our results suggest clear differences in results based on the instrument used, without corresponding student level data to better understand the relationship of teacher instructional practice to student outcomes, we cannot yet speak to the question of which instrument reflects teacher performance as it relates to student growth more accurately.

Conclusion

The primary goal of the RESET observation system is to detail practices at a level of specificity that, when used to provide an evaluation of a special education teachers' ability to implement evidence-based instruction, will give teachers a clear and consistent target for improvement. As Hill and Grossman (2013) note, "the absence of [specific] practices from most observation instruments limits the snapshot of teaching that emerges, the nature of feedback teachers receive, and the diagnostic information districts can glean about subject-specific needs for professional development" (p. 375). The results presented here add to the growing literature calling for observation instruments that are more aligned with the instructional practices relevant within specific content areas (Blazar et al., 2017; Hill & Grossman, 2013) and in particular for special education teachers, aligned with the evidence-based practices for students with disabilities (Johnson & Semmelroth, 2014; Jones & Brownell, 2014).

References

Blazar, D., Braslow, D., Charalambous, Y. C., & Hill, H. C. (2017). Attending to general and mathematics specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment, 22*(2), 71-94. https://doi.org/10.1080/10627197.2017.1309274

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. Educational and Psychological Measurement, 75(2), 311–337. https://doi.org/10.1177 /0013164414539163

Cash, A. H. Hamre B. K., Pianta R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542. https://doi.org/10.1016/j.ecresq.2011.12.006

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (2018). Variance and reliability in special educator observation rubrics. *Assessment for Effective Intervention*. https://doi.org/10.1177/1534508418781010

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Danielson, C. (2013). *The framework for teaching evaluation instrument.* Princeton, NJ: The Danielson Group. Deloitte, nd DATAUSA. https://datausa.io/profile/soc/special-education-teachers

Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt: Peter Lang.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom

observation protocol. *Teachers College Record*, *116*(6), 1-32. https://www.academia.edu/34327254/The_instructional_challenge_in_improving_teaching_quality_Lessons_from_a_classroom_observation_protocol

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*(6), 293-303. https://doi.org/10.3102/0013189X14544542

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055-2100. https://tedd.org/wp-content/uploads/2014/03/Grossman-et-al-Teaching-Practice-A-Cross-Professional-Perspective-copy.pdf

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1-28. http://scholar.harvard.edu/files/mkraft/files/herlihy_et_al._teacher_evaluation_systems_tcr.pdf

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction. *Cognition and Instruction, 26* (4), 430–511. https://doi.org/10.1080/07370000802177235

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203.

Hill, H., & Grossman P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83,*(2), 371-384. https://cepr.harvard.edu/files/cepr/files/ncte-hill-grossman-learning-from-teacher-observations.pdf

Holdheide, L., Warburton, S., & Buzick, H. (2012). State considerations in designing and implementing evaluation systems that include teachers of students with disabilities. In *Office of Special Education Programs Project Director's Conference*.

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2018a). Using evidence-centered design to create a special educator observation system. *Educational Measurement: Issues and Practice. 37*(2), 35-44. https://doi.org/10.1111/emip.12182

Johnson, E. S., Crawford, A. R., Moylan, L. A. & Zheng, Y. (2018b) Developing an explicit instruction observation rubric. *Journal of Special Education. 53*(1), 28-40. https://doi.org/10.1177/0022466918796224

Johnson, E. S., & Semmelroth, C. L. (2015). Validating an observation protocol to measure special education teacher effectiveness. *Journal of the American Academy of Special Education Professionals (online)*. https://files.eric.ed.gov/fulltext/EJ1134283.pdf

Johnson, E. S., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it matters and what makes it challenging. *Assessment for Effective Intervention. 39*(2), 71–82. https://doi.org/10.1177/1534508413513315

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2020, in press). The relationship of special education teacher performance on observation instruments with student outcomes. *Journal of Learning Disabilities*.

Jones, N. D. (2019, February). Observing special education teachers in high-stakes teacher evaluation systems. Presentation given at the Pacific Coast Research Conference, Coronado, CA.

Jones, N. D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention*, *39*(2), 112-124. https://doi.org/10.1177/1534508413514103

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Lawson, J. E. & Cruz, R. A (2018). Evaluating special educators' classroom performance: Does rater "type" matter?. *Assessment for Effective Intervention, 43*(4),1-14. https://doi.org/10.1177/1534508417736260

Linacre, J. M. (2017). *Facets 3.80* [Computer software].

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61-95.

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics, 9*, 1484–1509. http://dx.doi.org/10.1214/15-AOAS833

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater

errors. *Educational Measurement: issues and Practice, 34*(2), 34–46. https://doi.org/10.1111/emip.12061

McClellan, C., Donoghue, J., & Park, Y. S. (2013). Commonality and uniqueness in teaching practice observation. http://www.clowderconsulting.com/wp-content/uploads/2016/01/Commonality-and-Uniquenessin-Teaching-Practice-Observation_paper.pdf

McLeskey, J., Barringer, M. D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M.,& Ziegler, D. (2017). High-leverage practices in special education. *Arlington, VA: Council for Exceptional Children & CEEDAR Center. Google Scholar*.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom assessment scoring system (CLASS). Baltimore: Brookes.

Table 1

School Demographics of Participating Teachers

School	Grade	Enrollment (%Female)	White	Hispanic	Asian	Multi-race	Black	American Indian	%FRL	%SWD
1	K – 6	523 (48)	85	6	4	3	1	1	27	7
2	K – 5	470 (47)	75	14	4	4	2	1	54	9
3	6-8	1230 (50)	88	6	2	2	1	1	19	8
4	6-8	990 (49)	81	8	4	3	3	1	35	9
5	K– 5	729 (48)	76	9	7	4	3	1	63	11
6	K-6	358 (53)	79	11	4	4	1	1	63	7
7	K-5	664 (46)	52	45	1	1	1	1	68	10
8	K-7	810 (44)	88	6	4	2	1	1	16	4
9	K-6	368 (52)	79	7	5	5	2	2	98	10
10	K – 5	668 (49)	87	6	3	3	1	1	21	7
11	K-5	429 (44)	72	19	6	1	1	1	67	8
12	6-8	699 (44)	31	67	1	1	1	1	90	10
13	K-5	350 (51)	86	12	1	1	1	1	46	8
14	9-12	1369 (50)	87	9	1	1	1	1	34	8
15	K-6	511 (49)	59	27	2	1	9	1	100	10
16	K-5	498 (49)	28	70	1	2	1	1	95	8
17	K-6	518 (50)	89	4	3	1	2	1	31	9
18	K-8	359 (52)	85	8	2	3	1	1	16	5
19	6-8	906 (50)	63	32	1	1	1	1	64	8
20	6-8	711 (46)	41	55	1	2	1	1	87	7
21	K-6	163 (44)	91	4	1	3	1	1	53	9
22	K-5	643 (51)	65	31	1	2	1	1	64	8
23	K-3	292 (48)	69	21	4	2	3	1	40	9
24	4-8	345 (48)	70	19	1	2	7	1	49	8
25	K-12	60 (35)	90	9	0	0	1	0	33	45
26	K-5	508 (46)	65	31	2	1	1	1	40	8
27	K-8	252 (44)	88	5	5	2	1	1	38	5

Note. FRL = Free and Reduced Lunch, SWD = students with disabilities

Table 2

Teachers' overall fair average scores and rank order on RESET and FFT

Teacher Rank Order RESET	RESET score	FFT score	Teacher Rank Order FFT
7	2.79	2.68	7
51	2.66	2.22	54
6	2.63	2.21	57
2	2.56	2.11	16
11	2.50	2.19	34
26	2.49	1.99	51
23	2.46	2.07	6
32	2.45	2.08	11
34	2.42	2.26	14
24	2.41	2.06	2
54	2.39	2.44	9
17	2.38	1.74	32
30	2.37	2.07	23
14	2.29	2.15	30
4	2.25	1.80	24
16	2.24	2.34	58
57	2.23	2.38	31
19	2.21	1.58	15
15	2.16	2.00	26
5	2.11	1.88	5
9	2.11	2.11	41
58	2.09	2.02	10
41	2.06	1.88	4
31	2.04	2.02	8
10	1.90	1.81	17
8	1.85	1.77	52
52	1.79	1.68	19
Mean	2.29	2.06	
SD	.25	.25	

Table 3

Score Distributions Across Observation Instruments

Instrument	1	2	3
RESET	9%	51%	40%
FFT	17%	61%	23%

Table 4

Item correlations between RESET and FFT instruments.

	FFT Component 3a	FFT Component 3b	FFT Component 3c	FFT Component 3d	FFT Component 3e
RESET1	.35***	.01	.25*	.14	.18
RESET2	.32**	-.02	.26*	.14	.19
RESET3	.43***	.32**	.34***	.32**	.20
RESET4	.33***	.17	.14	.28**	.11
RESET5	.32**	.14	.12	.26*	.20
RESET6	.31**	.31**	.20	.36***	.09
RESET7	.30**	.20	.22*	.28**	.08
RESET8	.31**	.17	.21*	.21*	.15
RESET9	.37***	.22*	.27**	.17	.13
RESET10	.21*	.14	.17	.25*	.07
RESET11	.34***	.25*	.24**	.28**	.11
RESET12	.39***	.24*	.25**	.31**	.07
RESET13	.31**	.27**	.17	.28**	-.01
RESET14	.38***	.08	.12	.19	-.03
RESET15	.33***	.17	.11	.15	-.01
RESET16	.34***	.15	.14	.27**	-.05
RESET17	.19	.26**	.26**	.29**	.21*
RESET18	.38***	.09	.15	.19	-.01
RESET19	.22*	.19	.39***	.31**	.26*
RESET20	.42***	.19	.24**	.34***	-.04
RESET21	.34***	.33***	.39***	.45***	.12
RESET22	.22*	.39***	.48***	.29**	.18
RESET23	.17	.28**	.30**	.26*	-.03
RESET24	.22*	.31**	.30**	.38***	.15
RESET25	.28**	.28**	.29**	.23*	.23*

Note. RESET items are listed along the rows, and FFT items are listed along the columns. Cells with correlations at or above .30 are shaded. *p <.05, **p < .01, ***p < .001

Table 5

Items of RESET Explicit Instruction and FFT Domain 3

RESET Explicit Instruction Rubric Items
1. The goals of the lesson are clearly communicated to students.
2. The stated goal(s) is/are specific.
3. The teacher clearly explains the relevance of the stated goal to the students.
4. Instruction is completely aligned to the stated or implied goal.
5. All of the examples or materials selected are aligned to the stated or implied goal
6. Examples or materials selected are aligned to the instructional level of most or all of the students.
7. The teacher effectively reviews prior skills and/or engages background knowledge before beginning instruction.
8.The teacher provides clear demonstrations of proficient performance.
9. The teacher provides an adequate number of demonstrations given the nature and complexity of the skill or task.
10. The teacher uses language that is clear, precise, and accurate throughout the lesson.
11. Scaffolding is provided when it is needed to facilitate learning
12. Complex skills or strategies are broken down into logical instructional units to address cognitive overload, processing demands, or working memory
13. The teacher systematically withdraws support as the students move toward independent use of skills.
14.Guided practice is focused on the application of skills or strategies related to the stated or implied goal
15. The teacher consistently prompts students to apply skills or strategies throughout guided practice
16. The teacher maintains an appropriate pace throughout the lesson.
17. The teacher allows adequate time for students to think or respond throughout the lesson.

18. The teacher maintains focus on the stated or implied goal throughout the lesson.

19. The teacher provides frequent opportunities for students to engage or respond during the lesson.

20. There are structured and predictable instructional routines throughout the lesson.

21. The teacher monitors students to ensure they remain engaged.

22. The teacher consistently checks for understanding throughout the lesson.

23.The teacher provides timely feedback throughout the lesson.

24. Feedback is specific and informative throughout the lesson.

25. The teacher makes adjustments to instruction as needed based on the student responses.

FFT Domain 3 Components
3a. Communicating with students
3b. Using questioning and discussion techniques
3c. Engaging students in learning
3d. Using assessment in instruction
3e. Demonstrating flexibility and responsiveness