Hammill Institute on Disabilities

Journal of Learning Disabilities
2021, Vol. 54(1) 54–65

DOI: 10.1177/0022219420908906
journaloflearningdisabilities.sagepub.com
SAGE

The Relationship of Special Education Teacher Performance on Observation Instruments With Student Outcomes

Evelyn S. Johnson, EdD[1], Yuzhu Zheng, PhD[1], Angela R. Crawford, PhD[1], and Laura A. Moylan, MEd[1]

Abstract

In this study, we examined the relationship of special education teachers' performance on the Recognizing Effective Special Education Teachers (RESET) Explicit Instruction observation protocol with student growth on academic measures. Special education teachers provided video-recorded observations of three instructional lessons along with data from standardized, curriculum-based academic measures at the beginning, middle, and end of the school year for the students in the instructional group. Teachers' lessons were evaluated by external, trained raters. Data were analyzed using many-faceted Rasch measurement (MFRM), correlation, and multiple regression. Teacher performance on the overall protocol did not account for statistically significant variance in student growth beyond that of students' beginning of the year academic performance. Teacher performance on an abbreviated protocol comprised of items that had average or higher item difficulties on the MFRM analysis accounted for an additional 4.5% of variance beyond that of beginning of the year student performance. Implications for further research are discussed.

Keywords

rater accuracy, teacher observation, rater consistency, feedback, special education

Many students with learning and other disabilities (SWD) perform significantly below their peers in academic achievement on national- and state-level assessments (Gilmour, Fuchs, & Wehby, 2019; Schulte & Stevens, 2015). These large achievement gaps, estimated at 1.2 *SD* (Gilmour, Fuchs & Wehby, 2019), remain stable (Schulte et al., 2016) or worsen over time (Geary et al., 2012; Judge & Bell, 2010; Vaughn & Wanzek, 2014; Wei et al., 2011), suggesting that SWD are not receiving instruction that addresses their learning needs. Observational studies of instruction further support that SWD are not receiving effective instruction aligned with evidence-based practices (EBPs; e.g., Boardman et al., 2005; McLeskey & Billingsley, 2008; Vaughn et al., 2002).

Explicit instruction (EI) is an EBP supported by years of research (Hughes et al., 2017; Stockard et al., 2018) as an effective way to improve the achievement of SWD in both reading (Baker et al., 2006; Smolkowski & Gunn, 2012; Stockard et al., 2018) and math (Doabler et al., 2017; Gersten et al., 2009; Stockard et al., 2018). Despite this strong research base, observation studies of special education instructional practice suggest that EI may not be implemented on a large scale (Ciullo et al., 2016; McKenna et al., 2015; Swanson, 2008). This research to practice gap may explain why the achievement and growth of SWD continues to lag behind that of their peers without disabilities.

Teacher observation systems aligned with desired instructional practices offer one way to improve special education teachers' ability to implement EBPs, like EI, in the classroom and to address the achievement gap. Observation systems can promote teachers' instructional ability by identifying and defining effective practice, incentivizing its use, providing opportunities for feedback and informing professional development needs (Hill & Grossman, 2013; Papay, 2012). Emerging evidence supports the effectiveness of observation systems for improving instruction and students' academic outcomes (Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). However, limited evidence about the use of observation systems to improve *special education* instructional practices and outcomes for SWD exists.

[1]Boise State University, ID, USA

Corresponding Author:
Evelyn S. Johnson, Department of Early and Special Education, Boise State University, 1910 University Dr., MS 1725, Boise, ID 83725-1725, USA.
Email: evelynjohnson@boisestate.edu

Two issues explain why there remains such limited understanding of the impact of teacher observation on special education instructional practices and student outcomes. First, most of the commonly adopted observation systems were not designed with special education teachers in mind. If performance incentives are connected to the results of these observations, these systems may actively promote instructional practices that are not effective for SWD (Gilmour, Majeika, Sheaffer, & Wehby, 2019; E. S. Johnson, 2019; E. Johnson & Semmelroth, 2014; N. D. Jones & Brownell, 2014; N. Jones & Gilmour, 2019), and as a result, they are unlikely to lead to improvements in instructional practice for SWD. Second, it is difficult to connect performance on teacher observation tools with appropriate measures of student performance for SWD. Whereas student growth or achievement in many studies is operationalized as performance on standardized state assessments, it is challenging to link these results directly to teaching practices (Buzick & Weeks, 2018). In addition, some researchers have argued that distal measures like state assessments, may not be sensitive enough for SWD, nor adequately aligned with desired student outcomes (Fuchs & Fuchs, 2017).

RESET EI Observation Protocol

One promising tool to address the first issue is the Recognizing Effective Special Education Teachers (RESET) observation system (E. S. Johnson et al., 2018). RESET consists of 21 observation protocols aligned with EBPs for students with high incidence disabilities, including learning disabilities. The EI observation protocol is one of the 21 EBPs included within the RESET system, comprised of 25 items that detail the elements of EI (E. S. Johnson, Zheng, Crawford, & Moylan, 2019) and has been found to result in reliable evaluations of teacher practice across several studies (Crawford et al., 2019; E. S. Johnson, Zheng, et al., 2018, 2019).

Although the current evidence to support the reliability of evaluations is promising, no studies have yet been conducted to investigate the relationship of a special education teacher's ability to implement EI with their students' performance. EI has been determined to be an EBP, yet across studies, the detailed descriptions of EI vary, which limits our understanding of the *critical* elements that most impact student achievement (Durlak, 2010; Harn et al., 2013; E. S. Johnson, Zheng, et al., 2019). The RESET EI observation protocol details 25 items of this practice, which allows for the investigation of performance on specific elements of EI with student growth to begin to answer this question empirically. Item-response approaches such as Rasch measurement, allow for the identification of elements that help to distinguish between high, average, and low levels of implementation. Over time, these data also allow for the feedback and professional development support provided to special education teachers to focus on a smaller subset of elements as they work to improve their EI implementation.

Selecting Appropriate Measures of Student Performance

The second issue that makes testing the validity of special education observation systems challenging is the connection of teacher performance with appropriate measures of student academic achievement. Not only is it difficult to link state assessments directly to teaching practices (Buzick & Weeks, 2018), it is also difficult to ascertain the specific contribution of the special education instruction that a student may receive in addition to general classroom instruction. It has been argued that proximal measures, such as curriculum-based measures (CBMs) may be needed to determine the extent to which SWD are benefiting from instruction (Fuchs et al., 2018; Lynch et al., 2017). With the wide-scale adoption of multitiered systems of support (MTSS), many schools use CBMs on a regular basis to monitor progress in reading and mathematics. However, the students who are served by special education teachers may be working on a broad range of goals across content areas and grade levels. Therefore, aggregating growth across a variety of measures may be required to examine the relationship of teacher performance with student outcomes. One approach that has been used in other investigations of teacher observation and student performance is to transform test data to z scores (mean of 0, standard deviation of 1) to allow data to be aggregated across multiple test formats (see e.g., Borman & Kimball, 2005).

Understanding how quality of implementation of an EBP-like EI relates to student outcomes can provide practitioners with useful information not only about the level of implementation needed to realize improved achievement, but also about which elements of the practice may be the most important to focus on to achieve these outcomes (Harn et al., 2013). Therefore, the purpose of this study was to examine (a) the relationship between special education teachers' ability to implement EI as measured by the RESET EI observation protocol with student growth, and (b) whether a subset of EI elements can be identified that account for variance in students' academic growth.

Method

Participants

Special education teachers. Twenty-two special education teachers from 18 schools, and 7 districts from 3 states participated in this study. Teachers were recruited by sending study information and recruitment letters to the special education district directors, who shared recruiting materials with their special education teaching staff. All teachers were female. There was an even distribution of teachers across age levels, with four teachers between 20 and 29 years old, five teachers between 30 and 39 years old, six teachers between 40 and 49 years, five teachers were between 50 and 59 years old, and two teachers over 60 years old. One teacher was Hispanic, one was Asian and the remaining 20 were White. Their number of years' experience in special education ranged from 0 to 28 years (M = 7.88, SD = 8.33). Eighteen participants held a Bachelor's degree and four teachers held a Master degree in special education. Twenty teachers taught reading and two taught math. All participating teachers provided instruction in a resource room setting. Teachers reported using a variety of curriculum or teaching materials, including the *Wilson Reading System, Corrective Reading, Reading Horizons, Wonder Works, Attainment Math*, and *Do the Math*. Participating teachers received a stipend for providing video-recorded instruction of their classroom and de-identified academic information about their students.

Raters. Five male and 15 female raters were recruited from seven states in the United States to score videos in this study. Raters were recruited through a purposive sampling technique focused on selecting individuals who met the following criteria: held a teacher certificate, three or more years of experience in special education, and strong knowledge of EI (indicated by formal coursework or training and experience). A total of 17 raters were White, and 3 were Asian. All raters were special education professionals with between 3 to more than 20 years of working experience. One rater had a Bachelor's degree, 13 had a Master's Degree, and 6 had a Doctoral Degree. At the time of the study, seven raters worked as classroom special education teachers, six were in doctoral degree programs, five worked as a special education faculty or researcher at a university, one was a State Education Specialist, and one worked as an education curriculum developer. Eleven raters took formal coursework in EI when they were in an undergraduate or graduate program, two raters had additional experience supervising undergraduate students in an EI course as a Teaching Assistant in graduate school, and seven raters had in-service training in EI, or in-service training with a program that was designed using the principles of EI.

RESET EI Observation Protocol

The RESET EI observation protocol (see Appendix A) was used to evaluate participating teachers' ability to implement EI. This observation protocol consists of 25 items that detail the elements of EI (see E. S. Johnson et al. (2018) for a description of the RESET observation system development process). Each item is rated on a 3-point scale (1 = *Not implemented*, 2 = *Partially implemented*, and 3 = *Implemented*) to evaluate a teacher's level of proficiency in implementing that specific element. A number of studies have demonstrated that the EI protocol provides reliable assessments of a teacher's ability to implement this EBP (Crawford et al., 2019; E. S. Johnson et al., 2018, 2019).

Procedures

Video collection. All teachers provided video-recorded lessons of their instruction during the 2018–19 school year. Videos were recorded and uploaded using the Swivl™ capture system and ranged in length from 20 to 60 minutes. A total of 245 videos were provided, with the number of videos that each teacher contributed ranging from 8 to 16 (M = 11.14, SD = 1.86). Videos were organized into three categories (a) beginning of the school year (September–November), (b) middle of the school year (December–February), and (c) end of the school year (March–June). One video from each category for each teacher was randomly selected for analysis in this study for a total of three videos for each teacher, with 66 videos total.

Student data collection. Teachers provided progress monitoring data for each student who was in the instructional group that the teacher video recorded. The number of students per teacher ranged from 1 to 12 (M = 5.59, SD = 3.69), for a total of 117 students (see Table 1). Teachers were asked to provide beginning, middle, and end of year scores for each student using a standardized CBM. Seven different assessments were used by the teachers including eight iStation Overall Reading, six EasyCBM passage reading fluency, five AIMSweb passage reading fluency, two Star Reading, one AIMSweb Maze, one AIMSweb MCAP, one EasyCBM Numbers and Ops. Teachers were asked to indicate the student's grade-level placement and the grade level of the progress monitoring measures. Teachers were also asked to provide the raw score, the corresponding standard score (if applicable), and the corresponding percentile rank for the measures. Once all student data were received, the percentile scores were converted to z scores.

Rater training and scoring. Over a 4-day training period, raters were provided with an overview of the RESET project goals and a description of how the EI protocol was developed. Research project staff then explained each item of the EI protocol and clarified any questions the raters had. Raters were provided with a training manual that includes 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

Table 1. Summaries of Students' Data.

Teacher iD	Number of students	Student grade levels	Instruments
T1	12	2nd	iStation overall reading
T2	6	4th	AIMSweb fluency
T3	3	2nd, 4th, 8th	Easy CBM fluency
T4	4	4th, 5th	AIMSweb fluency
T5	3	4th	AIMSweb fluency
T6	3	3rd	iStation overall reading
T7	8	2nd	AIMSweb fluency
T8	10	2nd	iStation overall reading
T9	4	3rd	iStation overall reading
T10	7	3rd, 4th, 5th	Star reading
T11	3	1st, 3rd, 5th	Reading fluency
T12	2	6th, 8th	Easy CBM fluency
T13	10	8th	Easy CBM fluency
T14	2	4th, 6th	Easy CBM fluency
T15	5	2nd	iStation overall reading
T16	7	5th	Star Reading
T17	5	4th	iStation overall reading
T18	5	1st	iStation overall reading
T19	7	3rd	iStation overall reading
T20	1	3rd	AIMSweb maze
T21	5	5th	EasyCBM numbers and ops
T22	5	5th	AIMSweb MCAP
Total	117		

Note. CBM = curriculum-based measures; MCAP = Mathematics Concepts and Applications.

were reconciled with the master scored protocol. Disagreements in scores were reviewed and discussed. Raters were then assigned a randomly ordered list of videos and asked to evaluate the videos in the assigned order, to score each item, to provide time-stamped evidence 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 6 weeks to complete their ratings. Completed evaluations were submitted using an electronic version of the protocol developed in the Qualtrics® survey system.

To maintain a feasible observation load, we developed a rating scheme that allowed for scores across raters and videos to be linked without requiring each rater to score each video (Eckes, 2011). We randomly selected two teachers to have their first and last video scored by every rater. One rater was randomly selected to score at least one video of each teacher. Remaining videos were randomly assigned and each video was scored by three raters. This created a design in which all raters scored 25 videos.

Data Analysis

The scores assigned to the recorded lessons by raters were analyzed through many-faceted Rasch measurement (MFRM) analyses. The model used for the MFRM analysis in this study is given by:

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

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 difficulty 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 3.71 (Linacre, 2014). MFRM analysis produces infit and outfit statistics for each facet, two quality control statistics that indicate whether the measures have been confounded by construct-irrelevant factors (Eckes, 2011). Ranges in fit statistics from .5 to 1.5 are considered acceptable (Eckes, 2011; Englehard, 1992). FACETS also provides reliability and separation indices. The reliability index indicates the reproducibility of the measures if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007). Separation indicates the number of statistically distinguishable strata in the data. In addition, for the teacher facet, FACETS also computes a "fair average score," which result from a transformation of teachers' proficiency estimates reported in logits to the corresponding scores on the raw-scale score (Eckes, 2011). A fair average is the score that a particular teacher would have obtained from a rater of average severity and illustrates the effect of the model-based compensation for rater severity/leniency differences (Eckes, 2011).

Once the fair average score was computed for each teacher, we computed a median growth score for the students that each teacher worked with. Then, we calculated the Pearson's correlation coefficient between these two variables (teacher overall fair average score and median student growth). Finally, we conducted a multiple regression analysis using the students' growth scores as the outcome variable, and the teachers' EI protocol fair average score, student's grade level, and students' academic performance (in z scores) at the beginning of the school year as the predictor variables. These analyses were conducted using IBM SPSS, version 25 software. Using the results of the MFRM analysis, we then identified items with difficulty levels greater than 0 logits. Item difficulties are measured from the local origin of the item facet, where the average item has a difficulty of 0 logits (Linacre, 2017). We computed a new teacher fair average score based on these items

Measr	+Teacher	-Items	-Rater	-Lesson	Scale
4					(3)
3	T11				
		3			
2	T12				
	T1				
	T14				
	T19				-------
	T15				
1	T22, T9	13			
	T3, T8, T2				
	T6	25			
	T17	2	17		
	T7	7	11		
	T5	1, 24	12, 4		
	T10, T16, T18, T21	8, 9	1, 2, 5		
			16, 19	3	
0		12, 15, 16	18, 6	2	2
	T13	11	20, 8	1	
	T20	22	10, 13, 3		
		10, 14, 17, 18, 23, 4	14, 15		
		20, 6			
			7, 9		
	T4	21, 5			
-1		19			

-2					(1)

Figure 1. Variable map of the EI rubric facets items, teachers, raters, and lessons.
Note. EI = explicit instruction

and ran the correlation and regression analyses as described using the teacher fair average score on the abbreviated protocol.

Results

The results of the MFRM analysis are shown in Figure 1 and in Tables 2 and 3. Figure 1 includes the variable map and rank order of each facet. Tables 2 and 3 report the fit statistics and reliability and separation indices for the teacher and item facets, respectively. The far left column of Figure 1, titled "Measr," is the logit measure for the elements within each facet of the design. The second column contains the teacher measures, with more proficient teachers having larger logit values. Teacher 11 is the most proficient teacher, and Teacher 4 is the least proficient. The third

Table 2. Teacher Report From Many-Facet Rasch Measurement Analysis and Mean Student Growth.

Teacher number	Fair average	Median student growth (z scores)	Ability (logits)	Model *SE*	Infit MNSQ	Outfit MNSQ
11	2.89	−.27	3.05	.20	1.24	1.18
12	2.73	.52	2.02	.15	1.06	1.15
1	2.58	.58	1.47	.06	1.01	.97
14	2.54	.39	1.35	.11	1.01	1.05
19	2.50	.18	1.21	.11	1.11	1.11
15	2.48	.54	1.17	.11	.88	.88
22	2.43	.29	1.03	.12	.97	.92
9	2.42	.27	1.00	.11	1.27	1.25
3	2.40	.27	.93	.11	1.05	1.10
8	2.39	.11	.92	.11	.94	.94
2	2.38	.45	.89	.05	.96	.96
6	2.30	.13	.69	.10	.99	.88
17	2.29	.88	.66	.10	1.07	1.04
7	2.22	.11	.49	.11	.92	.92
5	2.15	−.59	.34	.10	1.28	1.28
16	2.12	.77	.27	.11	.93	.95
10	2.09	0	.21	.10	.75	.74
18	2.09	.55	.21	.11	.89	.89
21	2.09	.21	.20	.11	1.03	1.09
13	1.93	.78	−.16	.10	1.49	1.65
20	1.90	.74	−.22	.11	1.11	1.09
4	1.70	−.05	−.70	.11	.66	.66
M (count = 22)	2.30		.78	.11	1.01	1.02
SD	.29		.82	.02	.17	.18

Note. Root mean square error (model) = .12; adjusted *SD* = .81; separation = 6.98; reliability = .98; fixed chi-square = 1905.0; *df* = 47; significance = .00. MNSQ = mean-square.

column contains the item facet, with more difficult items having larger logit values, Items 3, 13, and 25 were the most difficult, and Items 19, 21, and 5 the least. As can be seen in the fourth column, raters differed somewhat in their severity levels, with raters 17, 11, 12, and 4 being less severe (more lenient), and raters 7 and 9 being more severe.

Table 2 reports the teachers' performance on the EI rubric, expressed in an overall fair average score, and includes the fit statistics and the reliability and separation indices for the teacher facet. The teacher's performance on the EI rubric ranges from a fair average score of 2.89 at 3.05 logits (*SE* = .20) for Teacher 11 who is the most proficient to 1.70, or .78 logits (*SE* = .11) for Teacher 4 who is the least proficient. The fit statistics fell within .66 to 1.49, which are within acceptable levels (Eckes, 2011). In addition to the fit statistics, reliability and separation information indices are reported. For teachers, the reliability coefficient was .98, and separation was 6.98, which demonstrate reliable differences in teacher proficiency.

Table 3 reports the fair average score of each item across teachers and lessons and includes the item difficulty expressed in logits, fit statistics, and the reliability and separation indices for the item facet. The item difficulty ranges from 2.24 logits (*SE* = .08) for Item 3 which is the most difficult, to −1.10 logits (*SE* = .09) for Item 19, which is the least difficult. For the item facet, the fit statistics fell within .78–1.56. The infit and outfit statistics for Item 3 of 1.56 and 1.63, respectively, fell outside the acceptable range of .50 to 1.50 (Eckes, 2011), suggesting unexpected patterns in the data. The reliability coefficient was .99, and separation was 8.71, which demonstrate reliable differences in item difficulty.

What Is the Relationship of Teacher Performance to Student Growth?

Student growth was quite variable across the sample of students (N = 117), ranging from −1.2 to 1.6 (M = .24, SD = .68). Teachers worked with a different number of students, so we first examined the correlation of teacher performance to student growth, by computing a median student growth score for each teacher (see Table 2). As can be seen, the median growth score ranges from a low of −.59 (for Teacher 5) to a high of .88 (Teacher 17), reflecting the variability in the student outcome data. The resulting coefficient was low and not statistically significant (r = .26, p = .24).

Table 3. Item Measure Report From Many-Facet Rasch Measurement Analysis.

Item number	Fair average	Difficulty (logits)	Model *SE*	Infit MNSQ	Outfit MNSQ
3	1.42	2.24	.08	1.56	1.63
13	1.93	.94	.07	1.16	1.16
25	1.93	.69	.07	1.13	1.13
2	2.09	.58	.07	1.19	1.17
7	2.11	.52	.07	1.26	1.25
1	2.18	.37	.07	1.16	1.15
24	2.18	.37	.07	.94	.99
8	2.25	.21	.07	.92	.98
9	2.25	.21	.07	1.00	1.04
12	2.33	.01	.07	1.04	1.06
16	2.34	.00	.07	.90	.92
15	2.34	−.01	.08	.92	.91
11	2.39	−.13	.08	.92	.96
22	2.42	−.23	.08	.90	.92
14	2.47	−.35	.08	.98	.96
10	2.47	−.36	.08	.88	.93
4	2.48	−.37	.08	.84	.89
23	2.49	−.41	.08	.88	.86
17	2.49	−.42	.08	.78	.82
18	2.49	−.42	.08	.86	.88
6	2.51	−.47	.08	1.06	1.11
20	2.52	−.49	.08	.96	.93
21	2.60	−.74	.08	.85	.90
5	2.60	−.74	.08	.87	.90
19	2.67	−1.01	.09	.86	.87
M (count = 25)	2.32	.00	.08	.99	1.01
SD	.27	.67	.00	.17	.17

Note. Root mean square error (model) = .08; adjusted *SD* = .67; separation = 8.71; reliability = .99; fixed chi-square = 1,703.3; *df* = 24; significance = .00. MNSQ = mean-square.

We conducted a stepwise regression analysis to determine the amount of variance in student growth accounted for by students' beginning of year score, their grade level, and teachers' overall fair average score on the EI observation protocol. The results of the stepwise regression analysis indicated that 12% of the variance in student growth was accounted for by the predictor variables, but only the students' achievement score from the beginning of the year was a significant predictor of the students' growth (see Table 4).

Next, we examined the results of the MFRM analysis to identify items with average or above item-level difficulties. As can be seen on the variable map in Figure 1 and in Table 3, 11 items of the EI protocol had item difficulties $>.00$. These include Items 1, 2, 3, 7, 8, 9, 12, 13, 16, 24, and 25. These items have consistently been shown to be the most difficult in previous studies with different teachers and raters using the EI observation protocol (see E. S. Johnson, et al., 2018, E. S. Johnson, Zheng, et al., 2019). We then computed a teachers' fair average score based on these 11 items (see Table 5). The ability levels on the abbreviated EI protocol ranges from 2.74 logits (SE = .27) for Teacher 11 who is the most proficient to −.80 logits (SE = .16) for Teacher 4 who is the least proficient. We also ran a Pearson's correlation coefficient between teacher performance on the abbreviated EI protocol items and median student growth, and found it was moderate but not statistically significant ($r = .34, p = .12$). Finally, we repeated the stepwise regression entering students' beginning of school year academic performance, grade level, and teacher fair average score on the abbreviated protocol (see Table 4). Together, the model accounted for 16% of the variance, with both beginning of the year student score (12%) and teacher performance on the abbreviated protocol (4.4%) accounting for unique variance in student growth.

Discussion

The purpose of this study was to examine the relationship of teachers' performance on an EI observation protocol with student outcomes and to determine whether specific elements of this practice may be identified as critical elements of EI. Our findings using the overall fair average score indicated a

Table 4. Multiple Regression Results for Predictors of Student Growth.

Variable	β	t	Significance	ΔR^2
Results with overall fair average score				
(Constant)		−1.558	.122	
Student beginning of year score	−.348	−4.016*	.001*	.121*
Student grade level (excluded)	.093	1.046	.298	
Teacher overall fair score (excluded)	.084	.956	.341	
Results with abbreviated protocol fair average score				
(Constant)		−2.748	.007*	
Student beginning of year score	−.383	−4.455	.001*	.121*
Teacher Fair Average Score	.213	2.473	.015*	.044*
Student grade level (excluded)	.105	1.208	.230	

*$p < .05$.

Table 5. Teacher Measure on Eleven Items Report From Many-Facet Rasch Measurement Analysis.

Teacher number	Fair average	Ability (logits)	Model *SE*	Infit MNSQ	Outfit MNSQ
11	2.86	2.74	.27	1.34	1.52
12	2.73	1.96	.23	1.09	1.27
1	2.45	1.02	.08	.98	.96
14	2.43	.98	.16	1.04	1.08
15	2.41	.94	.17	.84	.86
19	2.38	.85	.16	.95	.92
2	2.29	.64	.07	1.01	1.05
17	2.29	.63	.16	1.06	1.03
9	2.27	.59	.15	1.21	1.15
8	2.27	.59	.15	.93	.93
22	2.24	.52	.17	1.08	1.07
13	2.24	.52	.15	1.26	1.33
3	2.16	.36	.16	.86	.87
7	2.16	.35	.16	.84	.86
6	2.14	.31	.15	.77	.80
16	2.09	.19	.16	.99	1.01
21	2.02	.05	.17	1.21	1.37
5	1.99	−.02	.15	1.30	1.29
18	1.98	−.05	.16	.93	.92
20	1.84	−.34	.16	1.12	1.09
10	1.80	−.43	.16	.80	.79
4	1.64	−.80	.16	.71	.69
M (count = 22)	2.24	.59	.17	1.01	1.03
SD	.31	.83	.03	.17	.19

Note. Root mean square error (model) = .17; adjusted *SD* = .81; separation = 4.47; reliability = .96; fixed chi-square = 853.8; *df* = 47; significance = .00. MNSQ = mean-square.

very low correlation between a teacher's performance on the RESET EI protocol and median student growth, and also indicated that a teachers' overall fair average score did not account for variance in student growth above and beyond that accounted for by the students' beginning of school-year performance data.

One potential explanation for this finding is the variability in the student growth data. As seen in Table 2, median student growth ranged from a low of −.59 (median growth was negative for students taught by Teacher 5), to a high of .88 (median growth for students taught by Teacher 17 was the highest across the group). Median student growth was used to examine data across teachers serving a variable number of students. The variability in student growth across the entire sample of students ($N = 117$) was considerable, with a range of −1.12 to 1.6 ($M = .24$, $SD = .68$). With so much variability in student growth, it is not surprising that the predictor variables investigated in this analysis did not account for substantial variance in the dependent variable.

Rater leniency with regard to the teacher observation scores may also explain the lack of relationship between these two variables. In MFRM analysis, person abilities are measured from the local origins of all the other facets (Linacre, 2017). If average ability is high, then the average person has a positive logit measure. Examining the Wright map in Figure 1 shows that the overall performance of teachers on the EI observation protocol was high. As is shown in Table 2, 19 of 22 teachers had a positive logit measure. Upon further examination of the MFRM analysis, 51% of assigned scores across all items, teachers, lessons, and raters were a “3” for fully implemented, 32 % were a 2 for partially implemented, and 17% a 1 for not implemented.

Several large-scale studies and reviews of teacher observation and evaluation document what has been called the “Widget Effect” (Weisberg et al., 2009), where fewer than 1% of teachers in a district are rated as unsatisfactory, yet 81% of administrators and 57% of teachers can identify a teacher in their school as ineffective. A more recent study corroborates this finding, reporting that evaluators perceive more than three times as many teachers in their school as below proficient than they rate as such (Kraft & Gilmour, 2017). Although raters in our study were told that their evaluations of teacher performance would not be shared with the teachers, it is possible that they were unwilling to assign lower scores to teachers. Our data set does not allow us to examine this explanation further, but future studies examining differences in leniency between expert-scored protocols and those completed by trained raters could investigate whether trained raters also tend to be more lenient in their evaluation of teacher performance.

An additional, plausible explanation relates to the scale of the EI protocol. A 3-point scale does not provide a wide range with which to evaluate performance. The limited correlation could be due to a restriction of range, although the use of the fair average score allowed for the computation of a teacher evaluation score that was continuous. Extending the scale beyond the three descriptor levels also becomes problematic in practice, as the expanded range can negatively impact interrater reliability. In addition, to result in a valid evaluation, performance level descriptors must be based on a transparent evidentiary argument (Huff et al., 2010) and must allow for the reliable and meaningful discrimination across levels.

Finally, it is important to note that in this study, we only evaluated what Harn et al. (2013) have termed *process*-related evidence about a teacher’s ability to implement the EBP of EI but did not collect *structural* evidence such as the duration, frequency, intensity, or dosage per student (e.g., student attendance). Although dosage has been shown not to significantly impact student level outcomes in some studies of EBPs for SWD (Boardman, Buckley, Vaughn, Roberts, Scornavacco & Klingner, 2016), more pronounced discrepancies between research-based recommendations and practice may in fact, impact student outcomes and should be investigated in future studies.

To address the second study purpose, we identified critical elements of EI as those with difficulty levels at or above .00 logits. After calculating teacher evaluation scores based on an abbreviated score comprised of only items with higher difficulty levels, we did find a stronger, yet still statistically nonsignificant correlation between median student growth and teacher performance. We also found that teacher performance on these items accounted for a small, but statistically significant percentage of variance in students’ academic growth above and beyond that accounted for by their beginning of the school year performance.

A comparison of the items that had higher difficulty ratings with those with lower difficulty ratings does not produce a completely clear picture about what makes some items more difficult. However, the items with lower difficulty ratings tend to be features of published intervention programs such as alignment of the materials and examples to the goal (Items 4, 5, and 6), and the opportunity for guided practice, engagement and clear instructional routines (Items 14, 15, 19, and 20). Items with higher difficulty ratings tended to focus more on teacher-directed actions such as the statement and explanation of goals (Items 1, 2, and 3), the effective review of prior skills (Item 7), provision of clear and adequate demonstrations of proficient performance (Items 8 and 9), the ability to provide specific and informative feedback (Item 24) and to make adjustments as needed based on student response (Item 25). Interestingly, the 11 items identified in the current analysis as having higher difficulty ratings are consistently identified as such in previous studies conducted with the EI observation protocol (E. S. Johnson et al., 2018, 2019). This suggests that these items may be more useful in distinguishing teachers with more ability to implement EI effectively, and could provide a way to focus observation and feedback.

Finally, some researchers have suggested that when an evaluation instrument is content specific, teacher performance tends to be rated lower than when evaluated with a general protocol (Blazar, Braslow, Charalambous, & Hill, 2017). In our work developing RESET, we have seen a similar pattern in our results across studies. When testing the EI rubric, which focuses on the general aspects of EI, teacher ability has been centered between .8 and 1.2 logits. In studies to validate content-specific rubrics in reading and math, our results have reported teacher ability centered at −.35 logits (range −1.08 to .38) on a reading comprehension rubric (E. S. Johnson, Moylan, Crawford, & Zheng, 2019), and teacher ability centered at −.05 logits (range −1.73 to 1.37) on a math instruction rubric (Crawford et al., 2019). A direct comparison is not currently possible, as these studies involved different teachers and different raters, but further research investigating special education teachers' scores using an EI rubric as compared to a content-specific rubric in reading or math, and examining the relationship between teacher performance and student outcomes with the various teacher observation instruments could provide useful information about which protocols provide the most relevant information about a teacher's performance and the subsequent impact on student outcomes.

Limitations

In addition to the issues discussed, there are four limitations of the study. First, this study included a small number of teachers, who were all female, and predominantly White. This limits generalizability of the study's findings. Second, teachers' ability to implement EI was on three lessons. Research in teacher observation has shown that teacher performance can vary depending on the time of year and based on the students in class (Mantzicopoulos et al., 2018). However, several studies have reported that between 2 and 4 observations of teachers' instruction result in reliable estimates of teacher performance (Crawford et al., 2019; Kane & Staiger, 2012). Our findings found very little variability in teacher performance as a result of the lesson suggesting that this may not have been an issue in the current data set.

A third limitation is the variability in the length of the videos reviewed. The lessons ranged from 20 to 60 minutes, which represents a broad range across lesson time and introduces the concern of rater fatigue. To mitigate this concern, the assigned videos for each rater were varied in length (e.g., all raters scored some shorter and some longer videos), and the MFRM fit statistics and bias analyses did not indicate any consistent differences in scoring as a result of the lesson observed, suggesting that the length of the video may not have impacted raters' ability to provide consistent ratings across lessons. Finally, teachers' reported progress monitoring scores for their students, but we do not have information regarding the teachers' fidelity to standard administration or scoring of these measures.

Conclusion

The effectiveness research on EI provides a compelling rationale to support special education teachers to implement this EBP proficiently. The RESET EI observation protocol offers one way to help close the research to practice gap by providing teachers with reliable evaluations of their ability to implement this practice. Understanding how teacher evaluations are related to student growth can inform the identification of items that are strongly predictive and help focus attention to a smaller set of elements. The results in this study are encouraging in that we found that teacher performance on an abbreviated EI observation protocol did account for variance in student growth above and beyond that of a student's beginning of year performance. However, our results also suggest that the relationship is complex and that continued investigation of factors such as rater leniency and the structural aspects of EBPs are needed.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors.

References

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal*, *107*(2), 199–219.

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

Boardman, A. G., Argüelles, M. E., Vaughn, S., Hughes, M. T., & Klingner, J. (2005). Special education teachers' views of research-based practices. *The Journal of Special Education*, *39*(3), 168–180.

Boardman, A. G., Buckley, P., Vaughn, S., Roberts, G., Scornavacco, K., & Klingner, J. K. (2016). Relationship between implementation of Collaborative Strategic Reading and student outcomes for adolescents with disabilities. *Journal of Learning Disabilities*, *49*(6), 644–657.

Bond, T. G., & Fox, C. M. (2007). *Fundamental measurement in the human sciences*. Institute for Objective Measurement.

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *Consortium for Policy Research in Education Journal Articles*, *106*(1), 3–20. http://repository.upenn.edu/cpre_articles/2

Buzick, H., & Weeks, J. (2018). Trends in performance and growth by students with and without disabilities on five state summative assessments. *Applied Measurement in Education*, *31*(4), 269–282. https://doi.org/10.1080/08957347.2018.1495215

Ciullo, S., Lembke, E. S., Carlisle, A., Newman Thomas, C., Goodwin, M., & Judd, L. (2016). Implementation of evidence-based literacy practices in middle school response to intervention: An observation study. *Learning Disability Quarterly*, *39*(1), 44–57.

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (2019). Variance and reliability in special educator observation rubrics. *Assessment for Effective Intervention*, *45*(1), 27–37. https://doi.org/10.1177/1534508418781010

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297. https://doi.org/10.1002/pam.21818

Doabler, C. T., Clarke, B., Stoolmiller, M., Kosty, D. B., Fien, H., Smolkowski, K., & Baker, S. K. (2017). Explicit instructional interactions: Exploring the black box of a tier 2 mathematics intervention. *Remedial and Special Education*, *38*(2), 98–110.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "implementation research in early childhood education." *Early Childhood Research Quarterly*, *25*(3), 348–357. http://doi.org/10.1016/j.ecresq.2010.03.003

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

Englehard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171–191.

Fuchs, D., & Fuchs, L. S. (2017). Critique of the national evaluation of Response to Intervention: A case for simpler frameworks. *Exceptional Children*, *83*(3), 255–268.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology*, *104*(1), 206–223. https://doi.org/10.1037/a0025398

Gersten, R. M., Chard, D., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202–1242.

Gilmour, A. F., Fuchs, D., & Wehby, J. H. (2019). Are students with disabilities accessing the curriculum? A meta-analysis of the reading achievement gap between students with and without disabilities. *Exceptional Children*, *85*(3), 329–346. https://doi.org/10.1177/0014402918795830

Gilmour, A. F., Majeika, C. E., Shaeffer, A. W., & Wehby, J. H. (2019). The coverage of classroom management in teacher evaluation rubrics. *Teacher Education and Special Education*, *42*(2), 161–174. https://doi.org/10.1177/0888406418781918

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, *79*(2), 181–193.

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.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*(4), 310–324. https://doi.org/10.1080/08957347.2010.510956

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140–148.

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

Johnson, E. S. (2019, February). *Project RESET*. Paper presentation at the Pacific Coast Research Conference. Coronado, CA, United States.

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2018). 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., Moylan, L. A., Crawford, A., & Zheng, Y. (2019). Developing a comprehension instruction observation rubric for special education teachers. *Reading & Writing Quarterly*, *35*(2), 118–136. https://doi.org/10.1080/10573569.2018.1521319

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2019). Developing an explicit instruction special education teacher observation rubric. *The Journal of Special Education*, *53*(1), 28–40. https://doi.org/10.1177/0022466918796224

Jones, N., & Gilmour, A. (2019). Special education teacher evaluation: Examining current practices and research. In J. B. Crockett, B. Billingsley, & M. L. Boscardin, (Eds.), *Handbook of leadership and administration for special education* (2nd ed., pp. 458–477). Routledge.

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

Judge, S., & Bell, S. M. (2010). Reading achievement trajectories for students with learning disabilities during the elementary school years. *Reading & Writing Quarterly*, *27*(1–2), 153–178. https://doi.org/10.1080/10573569.2011.532722

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting *The Widget Effect*: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249. https://doi.org/10.3102/0013189X17718797

Linacre, J. M. (2014). *A user's guide to FACETS Rasch-model computer programs*. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=4A2EFB9FD14A09683984803B35314033?doi=10.1.1.177.9225&rep=rep1&type=pdf

Linacre, J. M. (2017). *Facets computer program for many-facet Rasch measurement, version 3.80.0*. Winsteps.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, *23*(1), 24–46.

McKenna, J. W., Shin, M., & Ciullo, S. (2015). Evaluating reading and mathematics instruction for students with learning disabilities: A synthesis of observation research. *Learning Disability Quarterly*, *38*(4), 195–207.

McLeskey, J., & Billingsley, B. S. (2008). How does the quality and stability of the teaching force influence the research-to-practice gap? A perspective on the teacher shortage in special education. *Remedial and Special Education*, *29*(5), 293–305.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141. https://doi.org/10.17763/haer.82.1.v40p0833345w6384

Schulte, A. C., & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children*, *81*(3), 370–387. https://doi.org/10.1177/0014402914563695

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. T. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test? *Journal of Educational Psychology*, *108*(7), 925–942. https://doi.org/10.1037/edu0000107

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *44*(2), 48–57.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of Direct Instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479–507.

Swanson, E. A. (2008). Observing reading instruction for students with learning disabilities: A synthesis. *Learning Disability Quarterly*, *31*(3), 115–133. https://doi.org/10.2307/25474643

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

Vaughn, S., Levy, S., Coleman, M., & Bos, C. S. (2002). Reading instruction for students with LD and EBD: A synthesis of observation studies. *The Journal of Special Education*, *36*(1), 2–13.

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research & Practice*, *29*(2), 46–53. https://doi.org/10.1111/ldrp.12031

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement of students with disabilities, ages 7 to 17. *Exceptional Children*, *78*(1), 89–106. https://doi.org/10.1177/001440291107800106

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009, June). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New Teacher Project. https://files.eric.ed.gov/fulltext/ED515656.pdf