Hammill Institute on Disabilities

Assessment for Effective Intervention
2023, Vol. 48(2) 113–123

DOI: 10.1177/15345084221112858
aei.sagepub.com
SAGE

Convergence Between Teacher Self-Report and School Administrator Observation Ratings Using the Classroom Strategies Assessment System

Nicole B. Wiggs, PhD[1], Linda A. Reddy, PhD[1], Ryan Kettler, PhD[1], Anh Hua, PhD[1], Christopher Dudek, MEd[1], Adam Lekwa, PhD[1], and Briana Bronstein, PhD[1]

Abstract

The Classroom Strategies Assessment System (CSAS) is a multi-rater, multi-method (direct observation and rating scale methodology) assessment of teachers' use of research-based instructional and behavior management strategies. The present study investigated the association between teacher self-report and school administrator ratings using the CSAS Teacher (CSAS-T) and Observer (CSAS-O) Forms in 15 high-poverty U.S. charter schools. The CSAS-T and CSAS-O were designed to be used concurrently as a valid formative assessment of teacher practice. Findings include small, but statistically significant correlations between the CSAS-T and CSAS-O. Analysis of a multi-trait–multi-method (MTMM) matrix found teachers and observers to be measuring different constructs. No mean score differences were found between teacher self-reported instruction and behavior management strategy use compared with school administrators' observed ratings. Furthermore, school administrators and teachers have similar ratings of overall effectiveness, with the majority of teachers in the sample being rated at or above effective. Overall, findings offer support for using the CSAS-O and CSAS-T for guiding professional development conversations.

Keywords

teacher assessment, evaluation, professional development, validity

Teachers' use of evidence-based strategies is necessary to promote student engagement with learning activities, as well as student academic achievement (Reddy, Hua et al., 2021). Teacher evaluation is designed to provide teachers with the formative feedback they need to improve classroom practices and addresses ongoing professional development (PD) needs of teachers (Reddy, Hua et al., 2021). Teacher evaluation is designed to monitor teacher performance and improve classroom practices by identifying teacher strengths and areas in need of improvement (Isore, 2009). Improving teacher evaluation has been a significant focus of education reform over the last 10 years, and the vast majority of U.S. states have implemented multi-method teacher evaluation systems that combine classroom observations with student achievement or growth measures. However, multiple threats to validity in teacher evaluation data exist, including bias in administrators' ratings of teachers (Kraft & Gilmour, 2017) and teacher dissatisfaction with and stress resulting from the evaluation process (Ford et al., 2017). Furthermore, student performance-focused evaluations do not account for classroom factors and processes of the teacher. Thus, teacher evaluation approaches that include multiple methods for assessing effectiveness and informing teacher PD improvements are needed (Cohen et al., 2014).

The use of student ratings and teacher self-report has emerged as appropriate, complementary sources of data for decision making post teacher evaluation (Nelson et al., 2016; Reddy, Dudek & Shernoff, 2016). Moreover, the use of teacher self-report has been emphasized by several national organizations (e.g., National Education Association and American Federation of Teachers) to improve the preparation for and delivery of instruction (Reddy, Dudek et al., 2016). Although a multi-method, multi-observer approach is advised as best practice in teacher evaluation, existing multi-rater methods often do

[1]Rutgers, The State University of New Jersey, Piscataway, USA

Corresponding Author:
Nicole B. Wiggs, Graduate School of Applied and Professional Psychology, Rutgers, The State University of New Jersey, 41 Gordon Road, Suite C, Piscataway, NJ 08854, USA.
Email: nb669@gsapp.rutgers.edu

Associate Editor: Lindsay Fallon

not emphasize the perspective of the classroom teacher in the process and instead heavily weight evaluation systems toward classroom observations.

Classroom Strategies Assessment System (CSAS)

The CSAS (Reddy & Dudek, 2014) was designed for school administrators and teachers to concurrently assess teachers' research-based practices. The CSAS is a multi-dimensional, multi-rater classroom assessment that consists of direct observation and rating scales using both observer and teacher perspectives. The CSAS was developed through a research-based process and contemporary test development theory (Reddy & Dudek, 2014; Reddy et al., 2013a, 2013b). The development process for CSAS dimensions and strategies was guided by input from effective experts, school stakeholder input, field testing across multiple pilot studies, and data analytic methods (see Reddy & Dudek, 2014; Reddy et al., 2013b).

The CSAS assesses evidence-based instruction and behavior management practices stemming from the literature on effective instruction and behavior management (e.g., Coutinho et al., 2005; Gable et al., 2009). Instructional and behavior management strategies present on the CSAS (see Table 1) are reflected in several seminal meta-analytic studies on effective instruction (Hattie, 1992, 2009; Marzano, 1998; Wang et al., 1993).

There are two forms of the CSAS that are designed to be used concurrently. The CSAS Observer Form (CSAS-O) can be used to observe teachers' practices, and the CSAS Teacher Form (CSAS-T) is a self-report tool that teachers can use to assess the same observed lesson. Using both forms is designed to promote dialogue about effective teaching practices and PD between observers and teachers (Reddy et al., 2013a, 2020).

Studies have demonstrated evidence of the predictive validity on the CSAS-O with students' academic proficiency and growth in achievement (Dudek et al., 2018; Reddy et al., 2013b, 2020), as well as student academic engagement (Lekwa et al., 2019). Research on each form individually evidenced acceptable construct validity and reliability (Reddy et al., 2013a, 2015), plus concurrent and convergent validity with other measures (Reddy et al., 2012, 2016, 2020). However, the relationship between the CSAS-O and CSAS-T, when used as part of the teacher evaluation process in low socioeconomic status (SES) school settings, has not been examined.

School Socioeconomic Status (SES) and Teacher Evaluation

Students residing in low-SES communities enter school at a disadvantage. Certain factors associated with living in low-SES areas (e.g., less exposure to books and entering school at a younger age) decrease school readiness (Janus & Duku, 2007). Poor school readiness can lead to lower academic achievement, higher drop-out rates, delinquency, and reduced mental health (Foster et al., 2005; Hair et al., 2006). The effects of low SES and school readiness are mitigated through high-quality instruction and interventions (Pears et al., 2014). Teacher evaluation using the CSAS can inform PD providing teachers in low-SES communities the skills to engage in high-quality instruction and behavior management (Reddy, Hua et al., 2021).

Purpose

The present study aims to evaluate the convergence between teacher self-report and school administrator ratings of classroom instructional and behavior management practices in urban schools serving low-SES communities using the CSAS.

First, we hypothesize small to moderate positive correlations between raters on the CSAS based on the timing of measurement (e.g., during instruction vs. teacher reflection) and previous research into inter-rater agreement on teacher evaluations. Reddy and colleagues (2013) found support for the hypothesis stating that teachers overestimate their ratings of instructional strategies in the classroom. Furthermore, Arnold-Berkovits et al. (2019) found strong, positive correlations between teacher self-reported time spent on instructional strategies (i.e., opportunity to learn) and observer ratings of corresponding instructional strategy use in classrooms. Given that the present study examined observers post-instruction ratings and not self-report estimates of time spent in instruction as measured by Arnold-Berkovits and colleagues (2019), we do not anticipate our results to yield as strong correlations. Instead, we predict differences in line with Reddy, Shernoff, and colleagues (2021), lowering the correlations to the small to medium range.

Second, we hypothesize that scores measuring the same construct using different methods were expected to have higher correlations than scores measuring different constructs within method. A multi-trait–multi-method (MTMM) matrix, which allows for the determination of variance between scores on parallel forms with multiple raters (Campbell & Fiske, 1959) is used to examine the convergent validity between ratings on the CSAS-T and CSAS-O. The MTMM is used to measure the cross-informant agreement across CSAS constructs (i.e., academic instruction and behavior management) using Cronbach's alpha to determine the greatest influence on scores, measurement type (i.e., rater), or trait (i.e., construct; Feeney-Kettler et al., 2019).

Finally, we expect variability between teacher self-report and school administrator ratings on use of evidence-based instruction and behavior management strategies in the classroom to exist, consistent with small to moderate

Table 1. CSAS-O and CSAS-T Strategy Rating Scales and Definitions.

Total, composite, or subscale	Definition
Total scores	
CSAS Total Score (IS + BMS)	Measures overall use of evidence-based instruction and behavior management strategies.
Instructional Strategies Total	Measures overall use of evidence-based instructional strategies.
Behavior Management Strategies Total	Measures the overall use of evidence-based behavior management strategies.
Instructional Strategies composites and subscales	
Instructional Methods Composite	Measures how classroom instruction occurs to attain how teachers incorporate learning in the classroom and how well academic content is delivered.
Adaptive Instruction	Strategies used to respond to student needs while teaching. Reflective of teacher flexibility and responsiveness to student needs and methods of differentiated instruction.
Student-Directed Instruction	How teachers engage students in the learning process, through constructivist and hands-on learning techniques.
Direct Instruction	Measures the strategies teachers use to deliver academic content and provide information to students.
Academic Monitoring and Feedback Composite	How teachers monitor student understanding of material and provision of feedback.
Promotes Student Thinking	The way teachers activate students' metacognitive and higher-order thinking abilities.
Academic Performance Feedback	How teachers provide feedback to students about their academic progress, including efforts to explain what is correct or incorrect about the students' academic performance.
Behavior Management Strategies composites and subscales	
Behavior Feedback Composite	The way teachers provide feedback about student behavior.
Praise	Verbal and nonverbal strategies used to provide praise to students for engaging in behavior that is appropriate for a classroom (e.g., following classroom rules, following directions).
Corrective Feedback	Verbal and nonverbal strategies used to correct or redirect students' inappropriate behaviors.
Preventative Methods Composite	How teachers promote positive and decrease negative behaviors in the classroom.
Proactive Methods	Verbal and nonverbal strategies used to increase engagement and decrease disruptive behaviors in the classroom. How teachers create a positive classroom environment.
Directives	How teachers communicate behavioral expectations and directions/instructions to students.

Note. CSAS-O = Classroom Strategies Assessment System–Observer Form; CSAS-T = Classroom Strategies Assessment System–Teacher Form; IS = Instructional Strategies; BMS = Behavioral Management Strategies.

strength correlations between raters. Given that teachers overestimate their ratings of instructional methods (Reddy & Dudek, 2016), we expect teacher ratings of evidence-based instruction and behavior management strategies higher than observers.

Method

Data for the current study were provided by the School Systems Improvement (SSI) Project, a large school reform grant sponsored by the U.S. Department of Education Teacher Incentive Fund (TIF; U.S. Department of Education, 2012). The TIF program supported local education agencies' (LEAs) development of comprehensive human capital management systems that included performance-based compensation systems for educators and rigorous educator evaluation systems based significantly on student growth. The TIF specifically focused on supporting high-needs LEA, defined as priority or persistently low achieving schools, and economically disadvantaged contexts (i.e., greater than 50% of students receiving free or reduced lunch percentage greater). Under the SSI Project, 15 economically disadvantaged charter schools in New Jersey (NJ) partnered with two universities to implement TIF program goals.

Participants

Data for the current study are derived from school administrator observations and teacher self-report scores on the CSAS that were required as part of a comprehensive teacher

evaluation system. The present sample consists of teachers in their first year of participation in the project. As per evaluation system requirements, all participants were observed three times over the course of the school year by their school administrator using the CSAS-O. Teachers completed three corresponding CSAS-T Forms, one for each corresponding CSAS-O observation.

Teachers. The sample consisted of 249 classroom teachers serving students in all grade levels (pre-kindergarten to 12th grade). Teacher ages ranged from 21 to 63 years old ($M = 32.5$; $SD = 9.2$) with the majority self-identifying as female ($n = 191$, 76.7%), White ($n = 179$, 71.9%), and Not Hispanic or Latinx ($n = 224$, 90.0%). Teachers also identified as Black or African American ($n = 38$, 15.3%), Asian ($n = 8$, 3.2%), Multi-Racial ($n = 6$, 2.4%), or Other ($n = 18$, 7.2%). The majority of teachers ($n = 184$, 73.9%) reported an earned bachelor's degree and 65 (26.1%) teachers reported an earned master's degree or higher.

School administrators. School administrators ($n = 30$) completed observations, with 70% identifying as female ($n = 21$). The administrators identify as African American ($n = 15$, 50%), White ($n = 12$, 40%), Asian ($n = 1$, 3%), and Other ($n = 2$, 7%). Highest degree earned also varied, 83% report an earned master's degree ($n = 25$) and five school administrators earned a doctorate (17%). Ages ranged from 30 to 54 years (M = 40.50, $SD = 7.95$). Observers reported an average of 15.43 years of teaching experience ($SD = 19.49$) and 7.25 years of administrative experience ($SD = 5.92$).

School context. The teachers and administrators in the current study come from 15 charter schools in the U.S. state of New Jersey participating in the SSI Project during the 2014–2015, 2015–2016, and 2016–2017 school years. All schools were located in urban or suburban areas across the northern, central, and southern regions of the state. Grade levels served, student enrollment, and student demographics varied for each school. Schools served elementary Grades K–5 ($n = 6$), elementary to middle school Grades K–8 ($n = 5$), Grades K–12 ($n = 2$), one school exclusively focused on high school Grades 9–12, and one school focused on Grades 6–12. Collectively all 15 schools represent 6,830 students across Grades K–12 ($M = 455$, $SD = 281$, range = 145–1,109). Schools serve a diverse body of students including Black/African American (40%), Hispanic/Latino (42%), Asian (6%), and White/European American (11.50%) among others. Students were evenly split 50.20% male to 49.80% female. Across schools, the average percentage of students receiving free reduced lunch was 72% ($SD = 0.16$, range = 51%–91%).

Measures

Classroom Strategies Assessment System. The CSAS is a multirater assessment that includes two forms, the CSAS-O for observers and the CSAS-T for teachers. Both forms examine the same areas of instruction and behavior management. Scores derived from each measure can be used to facilitate conversations about PD and effective strategies occurring in the observed lesson and classroom (see Table 1).

The CSAS-O and CSAS-T evidenced acceptable factor structure, test–retest reliability, inter-observer agreement, and internal consistency (Reddy et al., 2013a, 2015a). Both forms demonstrated evidence of predictive validity with student achievement for mathematics and ELA (English language arts) on statewide testing (e.g., Dudek et al., 2018; Reddy et al., 2013b, 2020) as well as students' academic engagement (Lekwa et al., 2019). Both forms were used as part of two randomized control trials on instructional coaching (Fabiano et al., 2018; Reddy, Hua et al., 2021).

Classroom Strategies Assessment System–Observer Form. The CSAS-O is completed by observers as part of a direct classroom observation process, and it includes three sections: (1) Strategy Counts, (2) Strategy Rating Scales, and (3) Classroom Checklist. During each observation, observers completed the Strategy Counts section by tallying the frequency of eight discrete teacher behaviors, while recording relevant evidence and observation notes. The Strategy Rating Scales were completed immediately after the observation and involved the review of Strategy Counts and observer notes. The Classroom Checklist was completed prior to the observer leaving the classroom, examining the presence or absence of important classroom procedures and physical artifacts.

The CSAS-O Strategy Rating Scales contain 54 items divided across two domains: Instructional Strategies (IS) and Behavioral Management Strategies (BMS). Items are summed together to create total, composite, and subscale-level scores. The IS rating scales contain 28 items that yield an IS total scale, two composite scales, and five subscales. The BMS rating scales contain 26 items that create a BMS total scale, two composite scales, and four subscales. The IS total and BMS total are summed together to create an overall CSAS total (IS + BMS) scale score. To complete the Strategy Rating Scales, observers first rated how often (*observed frequency rating*) teachers used specific strategies on a 7-point Likert-type scale (1 = *never used*, 3 = *sometimes used*, 7 = *always used*) and then rated how often the teachers should have used each strategy (*recommended frequency*) using the same Likert-type scale anchors. Recommended frequency scores are determined through several sources available to observers completing CSAS-O training, including the evidenced-based and effective teaching principles reviewed during training, scoring procedures and scoring mechanics of the CSAS

strategies, and alignment between the notes/evidence recorded during the observation procedure and the CSAS strategies.

A difference score, called the *discrepancy score*, is based on the absolute difference between the *observed frequency* and *recommended frequency* for each strategy. Discrepancy scores represent need for change with larger scores indicating a greater need for change. Discrepancy scores were calculated for each item and summed together to create subscale, composite, and total scale scores (see Table 1).

The CSAS-O includes guidelines for ratings of effectiveness (i.e., ineffective, partially effective, effective, and highly effective) based on the CSAS Total Discrepancy score. Administrators and teachers similarly matched the total score to the corresponding effectiveness rating guideline. Interpretation of this categorization reflects dichotomous decision making often used in schools. Either action is taken to address a teacher's effectiveness, or action is not taken.

Classroom Strategies Assessment System–Teacher Form. Teacher self-report ratings were captured using the CSAS-T. The CSAS-T includes the same Strategy Rating Scales and Classroom Checklist sections and items as the CSAS-O. The item scoring process and scale score calculations are also the same. Teachers first self-rated the *observed* and *recommended frequencies* for each item and then calculated a *discrepancy* score. The same corresponding effectiveness ratings guidelines are also found on the CSAS-T.

We examined convergence between the Strategy Rating Scales and subscale scores from the CSAS-O and CSAS-T. Interval-level scores and ratings of effectiveness were used for analysis.

Procedure

CSAS training. Trainings for the CSAS-O and CSAS-T were provided by individuals certified as CSAS trainers/master coders that successfully completed an extensive train-the-trainer model with CSAS authors. As a prerequisite, all CSAS trainers/master coders are required to complete the initial CSAS-O 3-day training (outlined below) for observers and demonstrate mastery in coding by exceeding a reliability threshold of 90% for the observer video-based criterion test. Trainers/mastercoders then successfully completed a 5-day training with CSAS authors reviewing the CSAS training protocols, video exemplars, scoring guides, and additional observer practice with classroom videos.

School administrator training. School administrators attended a 3-day training consisting of a four-step process familiarizing observers to the scientific literature, scoring, and utility of the CSAS-O. Step 1 oriented the observers to the scientific literature on effective academic instruction and classroom behavior management (e.g., Hattie, 2009) that guided the development of the CSAS-O dimensions and strategies. The second step involved training and practice on scoring procedures of the CSAS, which included (a) scoring procedures based on scientific literature, (b) CSAS scoring guidelines, and (c) alignment of school administrator's observation notes to the CSAS domains and subscales, emphasizing the observed and recommended frequencies. Step 3 required passing a knowledge test designed to ensure understanding and application of the training content. The fourth step involved in CSAS training required the school administrators to engage in video coding practice of classroom video exemplars and receive feedback from a CSAS trainer/master coder. To conclude training, the school administrators were required to pass a video-based criterion test using the CSAS and meet the minimum reliability score (70%). The video-based criterion test includes five classroom videos in which observers' scores are compared with master codes for each video to determine reliability scores. Master codes for each video were iteratively developed by CSAS authors and master coders who developed the CSAS. School administrators were provided additional in vivo, co-observation practice with a certified CSAS trainer/master coder.

Teacher training. The CSAS-T training is a two-step process aimed at enhancing teacher self-assessment beginning with a 3-hr instructional training session from a CSAS trainer. The CSAS theory, evidence, construct and item definitions, and criteria for scoring were discussed. The second portion was a discussion of the effective instruction literature and frequency of strategy use in relation to the CSAS. Teachers were not required to take a test to determine mastery. Instead, teachers reviewed and discussed classroom videos with the CSAS trainer/master coder.

Observation procedures. Teachers were observed a minimum of three times each year by their school administrator with the CSAS-O as part of routine performance evaluation requirements. Observations occurred throughout the academic school year during the fall, winter, and spring of each project year. As per teacher evaluation system requirements, the three CSAS-O observations were averaged together to create a summative end-of-year observation score for the CSAS-O. Teachers completed a CSAS-T for each observation conducted by their school administrator with the CSAS-O. This resulted in three CSAS-T observations for each teacher, which were then averaged together at the end of the year to create a summative self-reflection score for each teacher, which was required for the teacher evaluation process.

For the current study, the summative CSAS-O and CSAS-T score for each teacher was analyzed. Because CSAS-O administrator observations were paired with an accompanying CSAS-T self-reflection by teachers for each observation period prior to end-of-year summative procedures, the final summative end-of-year scores that

Table 2. Descriptive Data for the CSAS-O and CSAS-T Strategy Rating Discrepancy Scores.

	CSAS-O			CSAS-T		
CSAS Total, composite, or subscale	*M*	*SD*	Range	*M*	*SD*	Range
CSAS Total (IS + BMS)	26.40	20.20	0–103	29.8	23.9	0–132
IS Total	14.0	10.40	0–51.3	15.4	12.10	0–64
Instructional Methods Composite	7.21	6.08	0–29.7	8.52	6.92	0–32.3
Adaptive Instruction	1.81	1.70	0–9	1.93	1.64	0–7.3
Student DI	1.94	1.90	0–10.3	2.34	2.19	0–11.3
Direct Instruction	3.45	3.06	0–14	4.25	3.75	0–22.7
Monitoring and Feedback Composite	6.80	4.66	0–27.3	6.84	5.53	0–35
Promotes' Student thinking	2.87	2.35	0–12.3	3.20	2.75	0–14.3
Academic Performance Feedback	3.92	2.66	0–15	3.64	3.04	0–20.7
BMS Total	12.40	11.0	0–55	14.5	12.4	0–68.3
Preventative Methods Composite	4.60	4.58	0–28	7.44	6.91	0–37
Proactive Methods	3.17	3.16	0–14.3	3.85	3.66	0–22.3
Directives	2.82	2.85	0–14	3.59	3.48	0–17
Behavior Feedback Composite	6.46	5.96	0–31	7.01	5.89	0–31.3
Praise	3.48	3.05	0–16	3.20	2.81	0–19
Behavior Corrective Feedback	2.98	3.43	0–18.3	3.81	3.47	0–19

Note. CSAS-O = Classroom Strategies Assessment System–Observer Form; CSAS-T = Classroom Strategies Assessment System–Teacher Form; IS = Instructional Strategies; BMS = Behavioral Management Strategies; DI = Direct Instruction.

utilized an average of each form similarly maintained the pairing between the two forms.

Data Analyses

Relationship between teacher and school administrator CSAS ratings. Interval-level scores were used to address the first research question aimed at understanding the relationship between teacher self-report and school administrator ratings on the CSAS. *T* tests were conducted to compare any discrepancies between teacher self-report and observer ratings. Pearson correlations were used to evaluate convergence. Cohen's (1992) interpretation was used to analyze correlation strength. Correlations between .10 and .30 were considered small, .30 to .50 were considered moderate, and above .50 were considered large.

Effectiveness ratings. Descriptive analyses were used to analyze the second research question addressing teacher effectiveness. The CSAS Total score was used to estimate teachers' overall effectiveness, on a 4-point rating rubric (i.e., *ineffective, partially effective, effective*, and *highly effective*). Teacher and administrator categorical ratings were compared with each other and to guidelines provided by the CSAS. *T* tests were used to analyze score comparisons.

Results

Descriptive Analyses

Descriptive statistics of the CSAS-O and CSAS-T discrepancy scores for the current sample can be found in Table 2. On average, teacher self-report discrepancy scores were similar, but slightly higher than school administrator scores. For the CSAS Total scores, teachers self-reported (M = 29.8, SD = 23.9) an average discrepancy (i.e., need for change) score slightly higher than school administrator ratings (M = 26.4, SD = 20.2).

Relationship Between Teacher Self-Report and Observer Ratings

Cross-informant agreement on CSAS. As shown in Table 3, cross-informant correlations were found to be in the small range for CSAS-O and CSAS-T total, IS, and BMS total scores. When the CSAS measures were broken down by their composite scores, correlations of similar magnitude were found (rs = .19–.27). Correlations between subscales level also fell within the small range (r = .15–.27).

Within the sample, higher correlations were found between ratings on use of behavior management strategies (BMS Total r = .27) than ratings on the use of instructional practices (IS Total r = .19). The highest correlation was found between the CSAS-T and CSAS-O on the preventative methods composite (r = .27) indicating that ratings are more likely to converge on measures of positive reinforcement for behavior. Higher correlations were found between subscales measuring behavior management practices (e.g., behavior corrective feedback r = .26 and proactive methods r = .27) than academic instructional practices (e.g., adaptive instruction r = .15 and student-directed instruction r = .17). All correlations were significant at $p < .05$ (Table 4).

Table 3. Correlations Between CSAS-O and CSAS-T Strategy Rating Total, IS and BMS Total Scores.

		Teacher self-report		
Observer Report	Variable	IS	BMS	IS + BMS
Administrator observation	IS	**.19***	.17*	.18*
	BMS	.22*	**.27***	.25*
	CSAS Total (IS + BMS)	.22*	.23*	**.23***

		Teacher self-report			
		IM	MF	PreM	BF
Administrator observation	IM	**.20***	.13*	.14*	.12*
	MF	.19*	**.19***	.18*	.18*
	PreM	.24*	.18*	**.27***	.21*
	BF	.20*	.17*	.23*	**.26***

Note. Correlations between teacher scores are above the diagonal, correlations between administrator scores are below the diagonal, and teacher-administrator correlations are on the diagonal (bolded). CSAS-O = Classroom Strategies Assessment System–Observer Form; CSAS-T = Classroom Strategies Assessment System–Teacher Form; IS = Instructional Strategies; BMS = Behavioral Management Strategies; IM = Instructional Methods; MF = Monitoring and Feedback; PreM = Preventative Methods; BF = Behavior Feedback.
*Significant at $p < .05$.

Table 4. Correlations Between CSAS-O and CSAS-T Strategy Rating Subscale Scores.

	Teacher self-report								
Administrator observation	AI	SDI	DI	PST	APF	ProM	D	P	BCF
AI	**.15***	.18*	.14*	.10	.08	.09	.09	.08	.08
SDI	.15*	**.17***	.14*	.13*	.12	.09	.09	.09	.10
DI	.17*	.17*	**.19***	.14*	.11	.18*	.16*	.10*	.15*
PST	.16*	.19*	.20*	**.21***	.14*	.20*	.16*	.17*	.16*
APF	.12	.12	.16*	.13*	**.21***	.16*	.12	.19*	.13*
ProM	.22*	.22*	.23*	.20*	.16*	**.27***	.23*	.17*	.22*
D	.16*	.20*	.19*	.16	.14*	.23*	**.26***	.15*	.20*
P	.08	.10	.14*	.11	.15*	.17*	.15*	**.20***	.16*
BCF	.18*	.20*	.22*	.17*	.17*	.26*	.25*	.24*	**.26***

Note. Correlations between teacher scores are above the diagonal, correlations between administrator scores are below the diagonal, and teacher-administrator correlations are on the diagonal (bolded). CSAS-O = Classroom Strategies Assessment System–Observer Form; CSAS-T = Classroom Strategies Assessment System–Teacher Form; AI = Adaptive Instruction; SDI = Student-Directed Instruction; DI = Direct Instruction; PST = Promotes Student Thinking; APF = Academic Performance Feedback; ProM = Proactive Methods; D = Directives; P = Praise; BCF = Behavior Corrective Feedback.
*Significant at $p > .05$

Concurrent validity of the CSAS-T and CSAS-O. As shown in Table 5, a MTMM matrix was used to evaluate evidence for the concurrent validity of the CSAS-O and CSAS-T. The MTMM compared the measured traits of the CSAS across two raters using the BMS total and IS total. The four composite scores (IS, Monitoring and Feedback, Preventative Measures, and Behavioral Feedback) were included for a better understanding of convergent validity.

The MTMM analyses of the total scores and composite scores both indicated stronger shared variance within method (i.e., rater) than within trait (i.e., construct). Observer ratings are more highly correlated with observer ratings across constructs than ratings on the same construct by different raters. Correlations between IS and BMS were in the medium range for teacher raters ($r = .38$) and for observers ($r = .33$). Monotrait correlations for total scores were in the small range for both IS ($r = .19$) and BMS ($r = .27$). For composite scores, monomethod heterotrait correlations were in the large and very large ranges (rs =.64–.89), and heteromethod monotrait correlations were in the small range (rs = .19–.27). Heteromethod, heterotrait correlations were also in the small range and were mostly, slightly lower (rs = .13–.24).

Table 5. Multi-Trait, Multi-Method Correlation Matrix of CSAS Total and Composite Scores.

	Teacher–Observer		
CSAS composites	Monotrait *r*	Highest *r*	# higher *rs*
IS Total	.19	.38	2
Instructional Methods	.20	.89	4
Monitoring and Feedback	.19	.86	4
BMS Total	.27	.33	1
Preventative Methods	.27	.88	3
Feedback	.26	.77	3

Note. All correlations were significant at the $p < .05$ level. # higher *rs* = the number of Pearson correlations within the multi-trait, multi-method matrix (for the specified trait) that are larger in strength than the monotrait Pearson correlation. CSAS = Classroom Strategies Assessment System; IS = Instructional Strategies; BMS = Behavioral Management Strategies.

Comparisons of ratings conducted by school administrators and teachers. No group difference was found for total scores: Total ($t = 1.67$, $p = .090$), IS Total ($t = 1.35$, $p = .179$), and BMS Total ($t = 1.91$, $p = .057$). Teachers and school administrators rate similarly on teachers' need for change. Group differences between CSAS discrepancy scores were found for two of the four composites. Mean group differences on the instructional method ($t = 2.25$, $p = .025$) and preventive method ($t = 2.56$, $p = .011$) composites were found to be statistically significant. Discrepancy scores reported by teachers were significantly larger than scores reported by school administrators on both composite scores. No differences were found on the monitoring and feedback ($t = .096$, $p = .923$) and the behavior feedback ($t = 1.04$, $p = .300$).

Student-directed instruction, direct instruction, proactive measures, directives, and behavior corrective feedback were found to have statistically significant group differences. Teacher reported subscale discrepancy scores were significantly higher than school administrator observer ratings. Subscales with a statistically significant difference contribute to instructional methods and preventative methods composites.

Ratings of Effectiveness

Average teacher self-report ($n = 156$, 62.7%, $M = 29.81$, $SD = 23.90$) and school administrator ratings ($n = 173$, 69.4%, $M = 26.45$, $SD = 20.24$) fell in the *Effective* range, no difference was found between groups ($t = 1.70$, $p = .090$). Anecdotal differences were noted at the *Ineffective* level (administrator $n = 14$, 5.6%, and teacher $n = 25$, 10.0%). *Highly Effective* (administrator $n = 34$, 13.7%; teacher $n = 39$, 15.6%), and *Partially Effective* (administrator $n = 28$, 11.2%; teacher $n = 28$, 11.2%) ratings were largely commensurate.

Discussion

Teacher evaluation designed to identify and enhance teacher PD should include multiple sources and methods to capture the process of classroom teaching (ESSA, 2015). Using teacher self-report allows for reflection on the use of evidence-based classroom practices and can help improve the quality of classroom instruction (Desimone et al., 2010; Pollard et al., 2008). The present study examined the use of school administrator and teacher ratings of classroom instruction and behavior management practices across three observations. Findings revealed small positive correlations between raters. Overall, we found no differences between principal and teacher classroom practice ratings (i.e., CSAS Total, IS Total, and BMS Total).

Ratings of Classroom Instruction and Behavior Management Practices

Understanding small inter-rater correlations. Findings of the present study supported the hypothesis suggesting school administrator and teacher ratings of classroom practices would yield small to moderate correlations. Significant small positive correlations were found between ratings on all CSAS-O and CSAS-T scale scores. Findings are consistent with the magnitude of between-measures correlations in a previous predictive validity study that made use of both the CSAS-O and CSAS-T to predict student growth (Reddy, Hua et al., 2021) and with previous literature on multi-informant ratings of classroom observations and student outcomes (see Meyer et al., 2001; Reddy et al., 2013). One reason that may contribute to the small correlations between the two ratings is the nature of their administration. The CSAS-O ratings are generated from school administrators observing the teachers in the classroom and measuring the frequencies in real time. The CSAS-T ratings were completed after the lesson and require teachers to self-reflect. While reflecting, it is likely that teachers may have focused more on certain strategies they had hoped to improve on and were therefore, more critical of their own performance, which may have contributed to smaller correlations.

Given that the CSAS was designed to promote PD conversations between teachers and administrators, small correlations may be beneficial for facilitating PD conversations

between school administrators and teachers about what constitutes effective implementation. Differences between observed strategy use and self-reflected strategy use could guide discussions between teachers and observers with conversations centering on the highest rated need for change scores across the two raters and areas where differences exist between raters. For example, one of the instructional strategies measured by the CSAS is adaptive instruction, which includes efforts to differentiate instruction. Differentiated instruction in the larger literature has been defined as the way teachers respond to the differences in students' knowledge to amplify each student's potential (Tomlinson, 2000). However, the definition of differentiated instruction is not consistent across theory, professionals, and commercial packaging (Goddard & Kim, 2018). Adaptive instruction correlations could be constrained due to differing definitions for the measured constructs, despite both receiving training on the CSAS, or observer drift.

Convergent validity findings. Analyzed within an MTMM framework, evidence indicated that the CSAS-O and CSAS-T are measuring different constructs, or at a minimum measuring the same construct from distinct rater perspectives. All correlations between scores from teachers and scores from administrators on common traits were in the small range (rs = .19–.27), consistent with previous research on cross-informant agreement across a variety of studies, raters, and constructs (Meyer et al., 2001). Higher within-trait correlations are evidence of true construct measurement, rather than scores being reflective of irrelevant constructs such as rater bias or the halo effect. Although the results do not support convergent validity between observers and teachers, MTMM analyses further support the use of the CSAS-O and CSAS-T to enhance PD conversations.

Effectiveness of Classroom Teachers

The present research examined the difference between teacher self-reported and school administrator observed ratings of teacher effectiveness as measured by the CSAS. Findings suggest that school administrator and teacher ratings of effectiveness were largely consistent, falling into the *Effective* range. Average ratings for both groups also fell in the *Effective* range with no difference between groups.

Teacher ratings of effectiveness are often used in the evaluation process as a way to dichotomize scores into *effective* and *ineffective*. Using ratings in this way combines the various strategies being measured (i.e., instructional practices and behavior management strategies) into an "all or nothing" process. Teachers and administrators, alike, often overestimate the number of peers who should have ratings below proficient (Kraft & Gilmour, 2017). Thus, discussing differences in ratings of effectiveness may forge respect and meaningful PD conversations that advance opportunities for growth. Metrics of quantity are often missed in assessments of teachers' self-reported effectiveness (Herman et al., 2000; McCaffrey et al., 2001). Both CSAS forms include quantity and quality metrics for teachers' use of evidence-based practices, which could lead to ongoing performance feedback, support, and professional growth planning, which is often lacking in traditional PD models (e.g., Yoon et al., 2007).

Study Limitations and Future Direction

The present study is not without limitations. This investigation used data from a sample of school administrators and teachers employed in economically disadvantaged charter schools in NJ. Findings may not be generalizable across school administrators and teachers in the United States or to school personnel who do not teach in charter school or under-resourced settings. For example, the majority of school administrators in this sample were female and identified as African American, which may not be representative of school administrators in other states and regions. Likewise, the majority of teachers in our sample reported the highest degree earned as a bachelor's; however, 46% of charter school teachers in the United States reported having postbaccalaureate degrees (Tai & Goldring, 2020). Future research on multi-informant teacher evaluations should include a more representative sample from varied school settings from diverse locale reflecting national averages for race, gender, and degree.

All classroom observation data in the current study were collected by school administrators who were highly trained on the theory, evidence, and administration of the CSAS. These administrators received additional observation support throughout the school year and may have possessed unique skills in classroom observation assessment that may not parallel other school administrators using other teacher evaluation approaches. Generalizability in future research may improve by analyzing the convergence between school administrator observers and teacher self-reported ratings in routine evaluation, outside of a large grant-funded program.

Practical Significance of Findings

Findings support the use of teacher self-report in evaluation systems in schools located in low-SES areas as teachers present a complementary and unique view of their classroom practices. Furthermore, the significant correlations across raters offer support for the use of the CSAS to promote PD discussions between administrators and teachers. Post-observation and CSAS-T completion, teachers and administrator observers should discuss the classroom strategies used compared with recommended levels and allow for

teachers and school administrator observers to clarify discrepancies and determine teacher practice needs.

The CSAS conversations should lead to need-based teacher training. Introducing teacher PD associated with evidenced need could reduce burnout and turnover prevalent in low-SES school districts by improving teacher practices and knowledge of evidence-based instructional and behavior management strategies (Boyd et al., 2012; Reddy, Hua et al., 2021). Moreover, additional training of high-quality practice needs have been shown to improve student outcomes and mitigate the negative effects of economic factors associated with living in low-SES areas (Pears et al., 2014). Using the CSAS-T and CSAS-O can drive these necessary conversations and improve the quality of teacher–administrator conversations and increase necessary PD in schools.

Declaration of Conflicting Interests

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

Funding

The current study uses data from the School System Improvement (SSI) Project, a collaboration between multiple universities and charter schools funded by the U.S. Department of Education's Office of Innovation and Improvement as part of the Teacher Incentive Fund program (awarded to Rutgers, The State University of New Jersey; #S374A120060). The positions and opinions expressed in this article are solely those of the author.

ORCID iDs

Nicole B. Wiggs https://orcid.org/0000-0002-5417-5864
Linda A. Reddy https://orcid.org/0000-0001-8314-2810
Ryan Kettler https://orcid.org/0000-0003-4395-654X

References

Arnold-Berkovits, I., Kurz, A., & Reddy, L. A. (2019). Teacher log of students' opportunity to learn and classroom observation: An initial investigation of convergence. *Educational Assessment, Evaluation and Accountability*, *31*, 97–119. https://doi.org/10.1007/s11092-018-9288-2

Boyd, D., Grossman, P., Hammerness, K., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2012). Recruiting effective math teachers, evidence from New York City. *American Educational Research Journal*, *49*, 1008–1047. https://doi.org/10.3102/0002831211434579

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cohen, D. K., Peurach, D. J., Glazer, J. L., Gates, K. E., & Goldin, S. (2014). *Improvement by design: The promise of better schools*. University of Chicago Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Coutinho, S., Wiemer-Hastings, K., Skowronski, J. J., & Britt, M. A. (2005). Metacognition, need for cognition and use of explanations during ongoing learning and problem solving. *Learning and Individual Differences*, *15*, 321–337. https://doi.org/10.1016/j.lindif.2005.06.001

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, *24*(2), 267–329. https://doi.org/10.1177/0895904808330173

Dudek, C. M., Reddy, L. A., & Lekwa, A. (2018). Measuring teach practices to inform student achievement in high poverty schools: A predictive validity study. *Journal of Contemporary School Psychology*, *23*, 290–303. https://doi.org/10.1007/s40688-018-0196-8

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016). Retrieved from: https://www.ed.gov/essa?src=rn

Fabiano, G. A., Reddy, L. A., & Dudek, C. M. (2018). Teacher coaching supported by formative assessment for improving classroom practices. *School Psychology Quarterly*, *33*(2), 293–304. https://doi.org/10.1037/spq0000223

Feeney-Kettler, K.A., Kettler, R., & Dembitzer, L. (2009). Application of a multi-trait, multi-method matrix to a social, emotional, and behavioral secreening. *Journal of Applied School Psychology*, *35*, 20–36. https://doi.org/10.1080/15377903.2018.146884

Ford, T. G., Van Sickle, M. E., Clark, L. V., Fazio-Brunson, M., & Schween, D. C. (2017). Teacher self-efficacy, professional commitment, and high-stakes teacher evaluation policy in Louisiana. *Educational Policy*, *31*, 202–248. https://doi.org/10.1177/0895904815586855

Foster, S., Rollefson, M., Doksum, T., Noonan, D., Robinson, G., & Teich, J. (2005). *School mental health services in the United States, 2002–2003* [DHHS Pub. No. (SMA) 05-4068]. Rockville, MD: Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Gable, R. A., Hester, P. H., Rock, M. L., & Hughes, K. G. (2009). Back to basics: Rules, praise, ignoring and reprimands revisited. *Intervention in School and Clinic*, *44*, 195–205. https://doi.org/10.1177/1053451208328831

Goddard, Y. L., & Kim, M. (2018). Examining connections between teacher perceptions of collaboration, differentiated instruction, and teacher efficacy. *Teachers College Record*, *120*(1), 1–24. https://doi.org/10.1177/016146811812000102

Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., & Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly*, *21*, 431–454. https://doi.org/10.1016/j.ecresq.2006.09.005

Hattie, J. (1992). Measuring the effects of schooling. *Australian Journal of Education*, *36*(1), 5–13. https://doi.org/10.1177/000494419203600102

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement*. Routledge.

Herman, J., Klein, D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives.

Educational Measurement: Issues and Practice, *194*(4), 16–24. https://doi.org/10.1111/j.1745-3992.2000.tb00042.x

Isore, M. (2009). *Teacher evaluation: Current practices in OECD countries and a literature review* (OECD Education Working Papers No. 23). OECD. https://doi.org/10.1787/223283631428

Janus, M., & Duku, E. (2007). The school entry gap: Socioeconomic, family, and health factors associated with children's school readiness to learn. *Early Education and Development*, *18*, 375–403. https://doi.org/10.1080/10409280701610796a

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

Lekwa, A. J., Reddy, L. A., & Shernoff, E. S. (2019). Measuring teacher practices and student academic engagement: A convergent validity study. *School Psychology Quarterly*, *34*, 109–118. https://doi.org/10.1037/spq0000268

Marzano, R. J. (1998). *A theory-based meta-analysis of research on instruction* (ED427087). ERIC. https://eric.ed.gov/?id=ED427087

McCaffrey, D., Hamilton, K., Stecher, B., Klein, S., Bugliari, D., & Robyn, A. (2001). Interaction among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education*, *22*, 493–517. https://doi.org/10.2307/749803

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment. *American Psychologist*, *56*(2), 128–165. https://doi.org/10.1037/0003-066X.56.2.128

Nelson, P. M., Reddy, L. A., Dudek, C. M., & Lekwa, A. (2016). Observer and student ratings of the class environment: A preliminary investigation of convergence. *School Psychology Quarterly*, *32*(4), 465–479. https://doi.org/10.1037/spq0000176

Pears, K. C., Healey, C. V., Fisher, P. A., Braun, D., Gill, C., Conte, H. M., Newman, J., & Ticer, S. (2014). Immediate Effects of a Program to Promote School Readiness in Low-Income Children: Results of a Pilot Study. *Education & Treatment of Children*, *37*(3), 431–460. https://doi.org/10.1353/etc.2014.0021

Pollard, A., Anderson, J., Maddock, M., Swatfield, S., Warin, J., & Warwick, P. (2008). *Reflective teaching: Evidence informed professional practice* (3rd ed.). Continuum.

Reddy, L. A., & Dudek, C. M. (2014). Teacher progress monitoring of instructional and behavioral management practices: An evidence-based approach to improving classroom practices. *International Journal of School & Educational Psychology*, 2, 71–84. https://doi.org/10.1080/21683603.2013.876951

Reddy, L. A., Dudek, C. M., Rualo, A. J., & Fabiano, G. A. (2016). Concurrent validity of the Classroom Strategies Scale–Teacher form: A preliminary investigation. *Educational Assessment*, *21*, 267–277. https://doi.org/10.1080/10627197.2016.1236675

Reddy, L. A., Dudek, C. M., & Shernoff, E. S. (2016). Teacher formative assessment: The missing link in response to intervention. In S. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 607–623). Springer. https://doi.org/10.1007/978-1-4899-7568-3_34

Reddy, L. A., Fabiano, G., Dudek, C. M., & Hsu, L. (2013a). Development and construct validation of the Classroom Strategies Scale–Observer Form. *School Psychology Quarterly*, *28*, 317–341. https://doi.org/10.1037/spq0000043

Reddy, L. A., Fabiano, G., Dudek, C. M., & Hsu, L. (2013b). Predictive validity of the Classroom Strategies Scale–Observer Form on statewide testing. *School Psychology Quarterly*, *28*, 301–316. https://doi.org/10.1037/spq0000041

Reddy, L. A., Hua, A. N., Dudek, C. M., Kettler, R., Arnold-Berkovitz, I., Lekwa, A., Crouse, K., Kurz, A., & Hu, J. (2021). The relationship between school administrator and teacher ratings of classroom practices and student achievement in high-poverty schools. *Assessment for Effective Intervention*, *46*(2), 87–98. https://doi.org/10.1177/1534508419862863

Reddy, L. A., Kettler, R. J., & Kurz, A. (2015). School-wide educator evaluation for improving school capacity and student achievement in high-poverty schools: Year 1 of the school system improvement project. *Journal of Educational and Psychological Consultation*, *25*, 90–108. https://doi.org/10.1080/10474412.2014.929961

Reddy, L. A., Lekwa, A., Dudek, C., Kettler, R., & Hua, A. (2020). Evaluation of teacher practices and student achievement in high-poverty schools. *Journal of Psychoeducational Assessment*, *38*(7), 816–830. https://doi.org/10.1177/0734282920913394

Reddy, L. A., Shernoff, E., & Lekwa, A. (2021). A randomized controlled trial of instructional coaching in high-poverty schools: Examining teacher practices and student outcomes. *Journal of School Psychology*, *86*, 151–168. https://doi.org/10.1016/j.jsp.2021.04.001

Taie, S., & Goldring, R. (2020). Characteristics of Public and Private Elementary and Secondary School Teachers in the United States: Results from the 2017-18 National Teacher and Principal Survey. First Look. NCES 2020-142. National Center for Education Statistics.

Tomlinson, C. A. (2000). *Differentiation of instruction in the elementary grades* (ED443572 2000-08-00). ERIC Digest. https://www.eric.ed.gov

U.S. Department of Education. (2012). *Teacher incentive fund program*. https://oese.ed.gov/teacher-incentive-fund/

Wang, M. C., Haertal, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, *63*(3), 249–294.

Yoon, K. S., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007-No. 033). https://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/rel_2007033.pdf