Hammill Institute on Disabilities

Assessment for Effective Intervention
2023, Vol. 48(3) 149–158

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15345084221118316
aei.sagepub.com
SAGE

Measuring Classroom Management in Secondary Settings: Ongoing Validation of the Direct Behavior Rating-Classroom Management

Wesley A. Sims, PhD[1], Rondy Yu, PhD[1], Kathleen R. King, PhD[2], Danielle Zahn, BA[1], Nina Mandracchia, MA[1], Elissa Monteiro, MA[1], and Melissa Klaib, BA[1]

Abstract

Classroom management (CM) practices have a well-established, intuitive, and empirical connection with student academic, social, emotional, and behavioral outcomes. CM, defined as educator practices used to create supportive classroom environments, may be the implementation factor that is most impactful of the universal Tier I supports. Recognizing the importance of CM and existing deficiencies in pre- and in-service training for teachers, schools are increasingly turning to data-driven professional development activities as a solution. The current study continues the validation process of the Direct Behavior Rating-Classroom Management (DBR-CM), an efficient and flexible measure of teacher CM practices in secondary school settings. Data were collected from 140 U.S. Midwest middle and high school classrooms. Results found DBR-CM scores to be significantly correlated with several scores on concurrently completed measures of CM, including those that rely on systematic direct observation and rating scales. Findings continue the accumulation of validity evidence to address extrapolation, generalization, and theory-based inferences underlying the interpretation and intended uses of the DBR-CM. Results are promising and build on previous DBR-CM validation work. Limitations and implications are discussed.

Keywords

observation, professional development, progress monitoring

Classroom management (CM) practices are those employed by teachers to effectively create and maintain supportive and productive classroom environments (Back et al., 2016). An eclectic, multidimensional theory now guides contemporary conceptualizations of CM (Wallace et al., 2020). From this perspective, broadly, CM can be viewed as a set of practices, procedures, and behaviors that are used collectively and flexibly to facilitate desirable student outcomes. These individual practices, which are often associated with a specific theory (e.g., behaviorism, social learning), should be combined and flexibly utilized in specific contexts to create successful classroom environments. Examples of discrete CM practices encompassed within this perspective include the use of behavior-specific praise (Flora, 2000); opportunities to respond (Stichter et al., 2008); active instruction (Gage et al., 2018); precorrections (Reinke et al., 2015); engaging method of delivery of instructional content (Pianta et al., 2012); and efforts to create a safe, positive, and accepting classroom atmosphere (Back et al., 2016). Despite the well-established connection between CM practices and positive classroom atmosphere (Pianta et al., 2012) and student outcomes (e.g., student engagement, achievement, social competence; Back et al., 2016; Reinke et al., 2018), CM has garnered relatively little attention in both research and training (Christofferson & Sullivan, 2015; Cooper et al., 2018; Grasley-Boy et al., 2019). To mitigate the impact of this inattention, some scholars have called for the use of a data-driven, multi-tiered professional development (PD) model to promote

[1] University of California, Riverside, USA
[2] Palm Springs Unified School District, CA, USA

Corresponding Author:
Wesley A. Sims, University of California, Riverside, 900 University Dr., 1207 Sproul Hall, Riverside, CA 92521, USA.
Email: wesleys@ucr.edu

Associate Editor: Lindsay Fallon

effective CM practices (Simonsen et al., 2014; Sims et al., 2020). Unfortunately, the availability of defensible (i.e., psychometrically sound), flexible (i.e., varied applications, populations), and usable (i.e., acceptable, feasible) assessments of educator use of CM practices appears limited (Reddy et al., 2013). This study continues the accumulation of validity evidence to support use of the Direct Behavior Rating-Classroom Management (DBR-CM; Sims et al., 2020), an assessment of educator use of CM practices.

Deficiencies in Classroom Management Training

Educators consistently report feeling underprepared to promote academic engagement and, at the same time, preempt and reduce disruptive student behavior in their classrooms. As cited by teachers and confirmed by curricular audits, CM is largely neglected within pre-service teacher preparation (Christofferson & Sullivan, 2015; Cooper et al., 2018). In separate surveys, Christofferson and Sullivan (2015) and Cooper et al. (2018) found that only about half of the respondents received a dedicated course in CM, which typically focused only on antecedent-based strategies (e.g., establishing rules and classroom layout). Consistent with these findings, other appraisals of pre-service curricula for teachers (e.g., Begeny & Martens, 2006; Freeman et al., 2014) have found limited coverage of CM practices, leaving school districts and site administrators to address deficiencies in CM practice via in-service PD.

Professional Development in Classroom Management

Some scholars believe deficiencies in educator use of evidence-based CM practices may be best addressed through a multi-tiered approach to PD (Grasley-Boy et al., 2019; Simonsen et al., 2014). A multi-tiered system of educator support (MTSES; Sims et al., 2020) would utilize data to drive decision-making within a continuum of PD activities, like those commonly employed to alleviate challenges faced by students (e.g., Multi-Tiered Systems of Support, Positive Behavioral Interventions and Supports, Response to Intervention), to address a variety of skills and competencies, including CM. Within MTSES, ongoing and data-driven feedback-based PD activities are particularly salient.

The advantages of PD activities that emphasize skill use, like coaching and performance feedback (PF), are well documented. Such PD activities increase objectivity in information collection and use (i.e., reduce subjectivity; VanDerHeyden & Burns, 2018), add structure to processes (Noell et al., 2000), support reflection and improvement mechanisms (i.e., graphic summaries, trend analysis; Alvero et al., 2001), enhance overall effectiveness (Alvero et al., 2001; Noell et al., 2000), and are deemed acceptable and effective (Gage et al., 2017; Simonsen et al., 2017). Effective coaching and PF are predicated on the collection and use of data to support performance appraisal and inform collaborative discussions around strengths, areas for improvement, and methods for improving (Alvero et al., 2001; Reinke et al., 2009).

Classroom Management Assessment

Although research in CM assessment has increased as of late, additional work appears warranted given the limited number of usable CM assessment tools currently available. At present, many appear to lack appropriate psychometric defensibility or objectivity (Bracken & Fischel, 2006; Reddy et al., 2013) due to their reliance on inferences from student performance (Chetty et al., 2014), local rubrics, informal observation notes from principals and coaches, or teacher reports via questionnaires and checklists (Bracken & Fischel, 2006; Reddy et al., 2013). Furthermore, many available CM assessment tools have a limited focus as they were developed for use with specific interventions or attend only to a small number of behaviors of interest (e.g., praise; Bracken & Fischel, 2006; Reddy et al., 2013). In contrast, more defensible assessments may not be as feasible. Although systematic direct observation (SDO; see Reinke et al., 2015), rating scale (Reddy et al., 2013), and hybrid SDO-rating scale (Pianta et al., 2012) assessment methodologies are available, their use may be too costly and resource-intensive for use in some schools (Sims et al., 2020). The DBR-CM was developed to expand the available assessments of CM with which to generate data to support CM-focused PD activities within a MTSES framework.

DBR-CM Development

The four-step DBR-CM development process (see Sims et al., 2020) included: (a) a thorough literature review to identify discrete, evidence-based CM practices (e.g., behavior-specific praise, opportunities to respond), (b) content validation activities to create items based on similarity within CM practices identified in available literature (e.g., Praise, Communication), (c) developing operational definitions with examples and non-examples for each DBR-CM item, and (d) organizing items into a scoring format used by DBR Single Item Scales (Chafouleas, 2011). Item construction was grounded in a general outcome measure (GOM; Shinn & Shinn, 2002) assessment approach, in which several discrete skills or behaviors necessary for the successful completion of a skill or behavior are grouped by commonality (e.g., reading fluency, academic engagement) and assessed holistically to improve efficiency while retaining defensibility (i.e., reliability, validity; Shinn & Shinn, 2002). To this end, the *DBR-CM Praise* item included any teacher effort (e.g., verbal, tangible, physical) to positively reinforce contingent behavior. The *DBR-CM Communication* item was constructed to encompass any teacher effort to communicate expectations to students. The *DBR-CM Enthusiasm* item was selected to group teacher practices that facilitate student engagement in classroom activity and instruction. The *DBR-CM Rapport* item included teacher efforts to build and maintain positive relationships in their classrooms (see Table 1 in the online supplemental materials; see Sims et al., 2020).

Direct Behavior Rating-Single Item Scales

Research has established assessment using a structured DBR format as a defensible, flexible, and usable method for assessing behavior (Chafouleas, 2011). DBRs combine the strengths of GOMs (e.g., flexible, feasible, acceptable) and rating scales (i.e., psychometric defensibility) into a low inference assessment format that reduces the latency between observations and ratings, which is particularly well-suited for efficient collection of screening and formative assessment data (Chafouleas, 2011). Touted advantages of DBR use extend to (a) frugality of sustained attention and concentration to generate objective and reliable scores (Christ et al., 2009); (b) frugality of resources as it requires less training to use reliably (Harrison et al., 2014); and (c) intervention utility (e.g., PF, self-monitoring; Riley-Tillman, Chafouleas, Briesch, & Eckert, 2008).

Preliminary DBR-CM Validation Work

Consistent with Kane's (2013) argument-based validation model, assessment development begins with a clear statement of the intended interpretations and uses of generated scores and continues with the accumulation of validity evidence to support the proposed interpretation and use argument (IUA; Kane, 2013). The IUA for the DBR-CM proposed efficient generation of feasible, defensible, flexible, and usable observation data for screening and formative assessment of educator CM practices (Sims et al., 2020). Following development, validity evidence is accumulated across a network of five interrelated inferences that lead from instrumentation to score, then score to score interpretation and use. Scoring inferences relate to the ability of the assessment to generate a fair, accurate, and reliable translation of observed performance to a score. The generalization inference connects generated scores to expected performance (e.g., reliability, varied samples) and the extrapolation inference connects expected performance to performance in a variety of other related domains/outcomes (e.g., convergent validity, predictive validity). The decisional inference relates to applied decisions based on score interpretation (e.g., applied use). Theory-based inferences that link scores to latent variables or constructs that likely explain patterns in observed performance (e.g., connection to latent constructs, causal inference; Kane, 2013). Inferences are addressed through collection and dissemination of familiar psychometric reliability and validity evidence using samples representative of all intended assessment subjects.

Preliminary DBR-CM work in a sample of 107 elementary school classrooms found significant positive correlations between scores on the DBR-CM and concurrent measures of behaviors and characteristics of classroom instructional environments (e.g., Classroom Atmosphere Rating Scale, Brief Classroom Interaction Observation-Revised, Ohio Teacher Self-efficacy Scale; Sims et al., 2020). Significant correlations of varying sizes were also observed in the expected directions between DBR-CM item ratings and SDO variables (e.g., Brief Classroom Interaction Observation-Revised; Praise, Opportunities to Respond, Reprimands) of teacher CM behaviors. Similarly, in the absence of a formal DBR-CM training, measures of inter-rater reliability (IRR) exceeded acceptable levels for all but one item (i.e., Praise). While promising, the psychometric evidence provided in this initial study does not sufficiently support the use of the DBR-CM across all proposed interpretations and uses (Sims et al., 2020).

Current Study

This study continues the accumulation of validity evidence supporting the inferences underlying the IUA for the DBR-CM. Using new concurrent measures in a sample of secondary educators, findings build upon previously accumulated validity evidence. Four research hypotheses guided this study: First, acceptable levels of IRR were anticipated between DBR-CM ratings from multiple observers in a sample of secondary educators, which further addresses the DBR-CM generalization inference. The generalization inference in this context relates to the ability of the DBR-CM to consistently generate scores indicative of CM practice use across varied users, tasks, and contexts. Second, significant positive correlations were anticipated between the DBR-CM Total Score and concurrent measures of classroom atmosphere (i.e., Emotional Support, Classroom Organization, Instructional Support), which further support extrapolation and theory-based inferences underlying the DBR-CM. Third, significant correlations were anticipated between DBR-CM items and frequencies/rates of observed CM practices, which further support extrapolation and theory-based inferences. Specifically, significant positive correlations were anticipated between DBR-CM items and positive and proactive direct observation variables (e.g., praise, opportunities to respond), and significant negative correlations were anticipated between DBR-CM items and the use of reprimands. Last, significant positive correlations were anticipated between DBR-CM items and ratings on concurrent rating-based measures of educator efforts to develop and maintain supportive, productive, and warm classroom environments, which further support extrapolation and theory-based inferences. The extrapolation inference for the DBR-CM in this context extends generated scores to practical, real-world performance. This extension occurs through the examination of relationships between generated scores and existing measures (i.e., convergent validity). In addition, in this context, the theory-based inference for the DBR-CM relates to the generation of empirical evidence that aligns with or supports a multidimensional theory of CM, or a group of interrelated CM practices that collectively constitute good CM practice.

Method

Participants

Participants were 140 classroom teachers from nine middle and high schools in a U.S. Midwestern, mixed urban and suburban community. Demographic information was not reported by all participants. Of participants who reported demographic information, 80% (n = 56) identified as female and 20% (n = 14) identified as male. Racial/ethnic identities reported include 34.5% White, 12.4% Black, 1.1% Asian, and 0.6% Other; 51.4% did not provide a response. Years of teaching experience reported ranged from 1 to 23 (M = 10.4, SD = 6.3).

Measures

Demographic Data Questionnaire. A researcher-created, study-specific questionnaire was used to obtain general demographic information about participants in the fall and spring of each year of the larger efficacy trial. Solicited information included but was not limited to gender, ethnicity, education, grade level taught, and years of experience.

Brief Classroom Interaction Observation–Revised. The Brief Classroom Interaction Observation–Revised (BCIO-R; Reinke et al., 2015) is a behavioral observation system used to collect frequency/rate data on discrete teacher CM practices (e.g., use of behavior-specific praise). Frequency of targeted behavior occurrence was collected using the Multi-Option Observation System for Experimental Studies (MOOSES; Tapp, 2002) software on handheld computer devices. Acceptable levels of IRR (i.e., mean percent agreement = 88–90%), as well as significant intercorrelations among teacher behavior variables (r = .19–.36), are documented for the BCIO-R (Reinke et al., 2015).

Classroom Assessment Scoring System–Secondary. The Classroom Assessment Scoring System–Secondary (CLASS; Pianta et al., 2012) is an observational measure used to assess the quality of teacher-student interactions in middle and high school classrooms. The CLASS includes three domains of teacher-child interactions: Emotional Support, Classroom Organization, and Instructional Support. Items are rated on a 7-point scale from 1 = *low* to 7 = *high*. The CLASS is widely used in empirical research and has been shown to have moderate to high levels of IRR and significant intercorrelations among domains (Pianta et al., 2012).

Classroom Atmosphere Rating Scale. The Classroom Atmosphere Rating Scale (CARS; Wehby et al., 1993) is a measure of the overall quality of a classroom environment, with an emphasis on educator practices. The measure consists of 10 items that evaluate student levels of compliance during structured times and transitions, cooperation, adherence to rules, interest and involvement, on-task behavior, and the degree to which the classroom was supportive of student effort. Items are rated on a 5-point scale from 1 = *very low* to 5 = *very high*. The CARS has displayed acceptable IRR and high internal consistency, with alpha coefficients ranging from .94 to .95 (Wehby et al., 1993).

DBR-CM. The DBR-CM External Rater Form (DBR-CM ER; Sims et al., 2020) is an assessment tool used to evaluate teacher CM behavior. The measure specifically targets discrete classroom educator behaviors across four domains: Praise, Communication, Rapport, and Enthusiasm. Items are rated on an 11-point scale ranging from 0 = *low* to 10 = *high*, and ratings can be combined to compute a total CM score. The DBR-CM has been shown to have acceptable levels of IRR (e.g., intraclass correlation coefficient [ICC] values = .67 to .84) as well as significant associations with overall scores (i.e., BCIO-R, r = .41–.78; CARS, r = .41–.78; Ohio State Teacher Efficacy Scale, r = .25–.81) and individual scores (r = .41–.78) on concurrently completed measures (Sims et al., 2020).

Procedures

Study analyses were conducted using data collected at the final time point (i.e., Time 6) of a 2-year randomized control trial evaluating the efficacy of the CHAMPS program (see Sprick et al., 1998). Of the 70 original study participants, 42 (60%) were assigned to receive the study treatment. Participant intervention status was blinded from research personnel from a large research-intensive university in the Midwestern United States. Data collection activities were completed by approximately 20 observers (i.e., graduate researchers, principal investigators).

Observer training. Study personnel participated in training for all study measures. The BCIO-R training consisted of 90 min of in-person training, including 30 min of trainer-led practice with video clips, followed by live reliability checks in non-study classrooms for 2 weeks. All observers met the minimum criterion of 85% agreement with a BCIO-R master coder before collecting data. Similarly, CLASS training was conducted by a publisher-certified, on-site trainer and included 2 days of didactic instruction along with an extensive reliability check process (see https://teachstone.com/trainings). Trainings for the CARS and DBR-CM included respective 60-min presentations covering operational definitions of items and completion instructions but did not include practice observations or reliability checks.

Data collection. At the beginning of each year, participants were asked to complete the Demographic Data Questionnaire. Observation data were collected during 20-min observation sessions of classroom instructional periods. Each classroom observation was conducted with at least two observers per classroom and four observers for inter-observer agreement (IOA) observations. The combination of measures completed by observers was varied to limit possible priming or confounding effects on the completion of secondary measures. Observers first completed one of two primary outcome measures for the larger CHAMPS efficacy trial (i.e., BCIO-R or CLASS), then completed a secondary measure (i.e., DBR-CM or CARS). Thus, two assessment combinations existed: BCIO-R then DBR-CM and CLASS then CARS, or BCIO-R then CARS and CLASS then DBR-CM.

IOA data were collected for 32% of the observation-based measures used (i.e., BCIO-R, CLASS, DBR-CM, and CARS). If agreement fell below 90% on the BCIO-R or CLASS at any point during the data collection process, observers were required to complete additional training and reliability checks. Mean percent agreement for the BCIO-R was 92%. ICC values for the CLASS scores were acceptable (CLASS Overall = .95, Emotional Support = .94, Classroom Organization = .88, Instructional Support = .95). Similarly, the CARS ICC value was .90.

Data Analysis Plan

To evaluate levels of agreement between multiple raters, IRR for DBR-CM scores was evaluated using percent agreement and ICC values. Per recommendations from Riley-Tillman, Chafouleas, Sassu, et al. (2008), percent agreement for DBR-CM ratings was calculated at a ±1-point level. Ratings were considered in agreement if scores from two separate raters differed by no more than 1 point. Percent agreement values >.70 are considered acceptable (Lombard et al., 2002). ICC assesses IOA when behavior ratings are generated by multiple observers (Hallgren, 2012). ICC estimates can range between 1 and 0, with values closer to 1 indicating better agreement. Values <.40 are considered poor, values between .40 and .59 are considered fair, values between .60 and .74 are good, and values >.75 are considered excellent (Cicchetti, 1994). Based on prior IRR work with the DBR-CM, mean percent agreement values that approach or exceed the .70 threshold were anticipated.

To evaluate the remaining hypotheses, bivariate correlations were calculated between DBR-CM ratings and concurrently completed measures (i.e., BCIO-R, CARS, and CLASS). Coefficient values of .10 to .29 are considered small, .30 to .49 are medium, and .50 to 1.00 are large (Cohen, 1988). Rating anchoring for items was consistent across all study measures. Higher ratings or frequencies were considered more desirable, with individual exceptions for BCIO-R reprimands variables and CLASS Negative Climate scores. As a result, significant positive correlations of medium or higher magnitude were anticipated between DBR-CM scores and most scores on concurrent measures, with previously noted exceptions (i.e., use of reprimands).

Table 1. Interrater Reliability Information for DBR-CM Observations.

	% Agreement	Reliability		
DBR-CM Item/Score	±1	ICC	Lower	Upper
DBR-CM Praise	.47	.87	.64	.95
DBR-CM Communication	.76	.93	.81	.98
DBR-CM Enthusiasm	.83	.96	.90	.99
DBR-CM Rapport	.65	.90	.73	.96
DBR-CM Total	.64	.92	.88	.95

Note. ICC = intraclass correlation coefficient; DBR-CM = Direct Behavior Rating-Classroom Management.

Results

Descriptive data for DBR-CM, BCIO-R, CLASS, and CARS variables were calculated to examine overall distribution and variability across scores (see online supplemental Table 2). Observer ratings indicated generally homogenous and slightly more favorable levels for most measures and items, with scores clustering around, but generally above (or below for punitive, "negative" practices), the midpoint rating option of each scale.

The first hypothesis anticipated acceptable levels of IOA on DBR-CM ratings and was evaluated by calculating mean percent agreement and ICC values (see Table 1). Mean percent agreement met or exceeded the recommended .70 value for DBR-CM Communication (.73) and Enthusiasm (.83) items. Mean percent agreement failed to meet or exceed the recommended .70 value for DBR-CM Rapport (.65) and Praise (.47) items, as well as the DBR-CM Total Score (.64). Resulting ICC actual and confidence interval values fell in the good to excellent ranges, varying from .87 (Praise) to .96 (Enthusiasm).

Bivariate correlations were calculated to address the second hypothesis, which anticipated that DBR-CM Total Scores would be positively correlated with Overall or Composite scores on concurrently completed measures. Significant positive correlations were noted between all concurrently completed measures (see Table 2). Values ranged from .47 (BCIO-R Positive Implementation+) to .83 (CARS).

Table 2. Correlations Between DBR-CM Total and Overall Scores of Concurrent Measures.

Total/Overall Score	1	2	3	4	5	6
1 DBR-CM total	I					
2 Classroom Atmosphere Scale Total Score	.83	I				
3 BCIO-R Positive Implementation+	.47	.54	I			
4 CLASS emotional support	.75	.71	.46	I		
5 CLASS classroom organization	.63	.69	.38	.55	I	
6 CLASS instructional support	.67	.62	.56	.78	.51	I
7 CLASS student engagement	.65	.70	.33	.64	.76	.52

Note. All correlations were significant at the $p = .01$ level. DBR-CM = Direct Behavior Rating-Classroom Management; BCIO-R = Brief Classroom Interaction Observation–Revised; CLASS = Classroom Assessment Scoring System–Secondary.

Table 3. Correlations Between DBR-CM and BCIO-R.

	DBR-CM domain			
BCIO–R variable	Praise	Communication	Enthusiasm	Rapport
Rate of opportunities to respond	.08	.16	.25**	.18*
Rate of pre-corrections	.17*	.13	.04	.02
Rate of general praise	.56**	.20*	.27**	.35**
Rate of specific praise	.41**	.18*	.12	.09
Rate of total praise	.62**	.23**	.27**	.33**
Rate of explicit reprimands	−.30**	−.36**	−.30**	−.44**
Rate of harsh reprimands	−.10	−.10	−.18*	−.23**
Rate of total reprimands	−.31**	−.37**	−.31**	−.46**
Positive Implementation+	.39**	.36**	.40**	.44**

Note. DBR-CM = Direct Behavior Rating-Classroom Management; BCIO-R = Brief Classroom Interaction Observation–Revised.
*$p = .05$. ** $p = .01$.

The third hypothesis anticipated significant correlations in expected directions (i.e., positive for all variables *except* for reprimand variables) between individual DBR-CM items and concurrent direct observation (BCIO-R) variables. As anticipated, several significant correlations were evident between DBR-CM items and rates of observed classroom behaviors (see Table 3). Overall, correlations between individual DBR-CM items and Rates of Precorrection and Opportunities to Respond were relatively unremarkable. Significant correlations were seen between DBR-CM items and BCIO-R variables measuring teacher praise and reprimand use. Finally, as anticipated, significant correlations were evident between DBR-CM items and the BCIO-R Positive Implementation+ variable, a measure of overall CM practice use.

The fourth hypothesis anticipated significant correlations between DBR-CM items and concurrent CARS and CLASS items. Significant positive correlations were noted between all DBR-CM and CARS items (see Table 4). All correlations between DBR-CM and CARS items were positive and significant, but some were more noteworthy than others. Significant, large, and positive correlations were observed between DBR-CM Communication, Enthusiasm, and Rapport items and all CARS items. In contrast, a large, positive, and significant correlation was only noted between the DBR-CM Praise and CARS Supportive items ($r = .70$, $p = .01$). Correlations between DBR-CM Praise and the remaining six CARS items were significant, medium, and positive. Like the CARS, correlations between all CLASS and DBR-CM ratings were significant and in expected directions (see Table 5). Again, some appeared larger in magnitude and more remarkable than others. Significant large correlations were noted between the DBR-CM Praise and CLASS Positive Climate and Negative Climate ratings. The DBR-CM Communication item appeared most largely correlated with the CLASS Teacher Sensitivity, Behavior Management, Productivity, Negative Climate Instructional Learning Formats, and Content Understanding scores. Large significant correlations were noted between the DBR-CM Enthusiasm ratings and all CLASS items except the Analysis and Inquiry item. Large significant correlations were found between the DBR-CM Rapport item and the CLASS Positive Climate, Teacher Sensitivity, Regards Adolescent Perspective, Behavior Management, Productivity, Negative Climate, Instructional Learning Formats, and Quality of Feedback items.

Table 4. Correlations Between DBR-CM and CARS.

	DBR-CM			
CARS items	Praise	Communication	Enthusiasm	Rapport
Compliance	.45	.62	.61	.68
Rules	.43	.58	.57	.65
Cooperation	.33	.59	.56	.64
Interest	.40	.57	.73	.67
Focused	.42	.70	.66	.68
Individual differences	.45	.65	.58	.67
Supportive	.70	.54	.63	.71

Note. All correlations were significant at the p = .01 level. DBR-CM = Direct Behavior Rating-Classroom Management; CARS = Classroom Atmosphere Rating Scale.

Table 5. Correlations Between DBR-CM and CLASS.

	DBR-CM			
CLASS items	Praise	Communication	Enthusiasm	Rapport
Positive climate	.61	.48	.61	.72
Teacher sensitivity	.46	.55	.60	.71
Regards adolescent perspective	.37	.39	.51	.51
Behavior management	.47	.66	.58	.72
Productivity	.45	.66	.59	.65
Negative climate	−.51	−.54	−.54	−.69
Instructional learning formats	.36	.58	.61	.59
Content understanding	.27	.56	.59	.49
Analysis and inquiry	.31	.38	.47	.40
Quality of feedback	.46	.41	.53	.56
Instructional dialogue	.34	.41	.55	.43

Note. All values significant at the p = .01 level. DBR-CM = Direct Behavior Rating-Classroom Management.

Discussion

Given their unequivocal link, it is not surprising that poorer student outcomes and teacher attrition may be attributable to teacher-reported deficiencies in CM training (Back et al., 2016). These reported deficiencies highlight the importance of PD activities emphasizing skill development and use through regular coaching and PF to improve CM practices (Simonsen et al., 2017). The efficient collection of flexible and defensible (i.e., reliable, valid) data is essential to such approaches to PD. This study continued the accumulation of validity evidence in a previously unexamined sample to support the generalization, extrapolation, and theory-based inferences underlying the IUA for the DBR-CM for use within such PD applications.

Interobserver Agreement

Consistent with the first hypothesis guiding this study and prior results found in a sample of elementary educators (Sims et al., 2020), IRR values exceeded desired levels for several DBR-CM items (i.e., Communication, Enthusiasm, and Rapport). Unfortunately, values for reliability metrics fell below desired levels for the DBR-CM Praise item. The emphasis on discrete instances of praise by the BCIO-R measure may explain these less than desirable results for the DBR-CM Praise item. In other words, the difference in measure scoring formats, observers' tallying of actual use of praise (i.e., BCIO-R), and the Likert scale format of the DBR-CM may have resulted in inconsistencies in the rating of praise on the DBR-CM. The resulting ICC values indicated higher levels of IRR for all DBR-CM items when compared with percent agreement results. Given ICC estimates the magnitude of the relationship between ratings overall while accounting for potential rater bias, resulting values present a more favorable picture of the ability of raters to use the DBR-CM reliably. Overall, when considering the cursory DBR-CM training provided to study observers, positive findings may reflect the high feasibility and usability typically associated with the DBR assessment methodology (Chafouleas, 2011; Riley-Tillman, Chafouleas, Briesch, & Eckert, 2008). Furthermore, results may have been adversely impacted by the brief observer training provided and the influence of concurrent measure completion relative to DBR-CM completion. It is reasonable to believe that additional more thorough training and reliability checks, as well as the use of additional personnel during observations to limit potential cross-measure confounding influences, would increase IRR values to desired levels (see Schlientz et al., 2009).

Generally, these findings support the accumulation of additional validity evidence supporting the generalization inference for the DBR-CM IUA, or to reliably generate scores indicative of CM practice use across varied users, tasks, and contexts (Kane, 2013). Findings were generated in a novel sample of secondary educators, collected by a unique group of observers, in a secondary setting (i.e., context), and during secondary-level instructional activities (i.e., task; Kane, 2013). Furthermore, findings support generalization inferences underlying the DBR-CM IUA by indicating a reasonable likelihood of obtaining reliable scores across novel observers and target subjects.

Associations Between Concurrently Completed Measures

Consistent with the remaining study hypotheses, as anticipated, noteworthy agreement was evident between DBR-CM scores and scores generated from concurrently completed measures of CM practices. First, associations between the DBR-CM Total Score and CARS Total Score, the BCIO-R Positive Implementation+ variable, CLASS Total score, and CLASS Domain scores (i.e., Emotional Support, Classroom Organization, Instructional Support) provide evidence of convergent validity. These findings suggest the DBR-CM measures CM as a broad concept (i.e., total score) in a manner consistent with similar measures of CM. Although favorable generally, relationships noted between the DBR-CM and CLASS are of particular note. The CLASS is considered a "gold standard" measure of CM, and these results provide excellent convergent validity evidence in support of the DBR-CM. These findings continue the collection of validity evidence addressing the extrapolation and theory-based inferences for the DBR-CM. Study results support the extrapolation inference by providing evidence that connects generated DBR-CM scores to practical, real-world performance (i.e., convergent validity). Evidence of convergent validity at the total score level also addresses the theory-based inference within the DBR-CM IUA. These findings support a multidimensional theory of CM, where a variety of elements or behaviors are subsumed by a broad, holistic concept or construct (i.e., CM; Wallace et al., 2020). Additional evidence of convergent validity was noted at the item level via significant associations between individual DBR-CM items and CARS items, BCIO-R variables, and CLASS items measuring positive CM practices. As with associations noted between total or domain scores, identified significant relationships connecting DBR-CM items to existing measures of targeted domains (Kane, 2013) address the extrapolation inference within the DBR-CM IUA. These results also further extend the evidence addressing the theory-based inference for the DBR-CM IUA. Findings suggest that DBR-CM items measure discrete yet interrelated CM practices that collectively constitute good CM practice in a manner consistent with existing measures. Unfortunately, the multidimensional, interrelated nature of CM may make distinguishing unique elements of CM difficult to discern at the measurement level. However, from an operational definition perspective, discrete CM practices are clearly unique. For example, providing opportunities to respond (i.e., facilitating engagement in instruction) and non-contingent reinforcement (i.e., building rapport) are qualitatively different but are associated with desirable CM practices individually or collectively. Ultimately, it is important to note that assessments should be selected while considering their intended use, including the implications of their use (Riley-Tillman et al., 2005). The DBR-CM is grounded in a GOM assessment approach that may exacerbate the apparent interconnectedness of generated scores. Users should consider the benefits and challenges associated with the DBR-CM when selecting an assessment to support their intended uses.

Limitations and Future Directions

Although findings are promising, this study is not without limitations. First, the limited sample of participants from a specific geographic area is not representative of all educators everywhere and may in part explain the limited variability in observed CM practices. Future research should accumulate additional validity evidence across samples that vary in terms of location, race/ethnicity, grade level, and other demographic traits. Second, as previously noted, the training provided to observers on the use of the DBR-CM was limited to a brief verbal explanation of items and written instructions on the rating form. Although IRR was found to be generally acceptable, future research should include more thorough, formal DBR-CM training with reliability checks for users and should explicitly explore the impact of training on accuracy and reliability. Third, raters completed multiple measures for each observation. It is possible that this arrangement may have produced unintended sequencing or order effects. In future studies, separate observers should be assigned to complete each concurrent measure. Lastly, to address potential limitations noted in the interrelatedness of items or CM variables, future efforts should focus on the accumulation of additional validity evidence that attempts to better differentiate unique dimensions of CM. Similarly, future research should begin the accumulation of validity evidence to address the decisional or use inference within the DBR-CM IUA. The former would endeavor to identify measures, items, and variables (i.e., behaviors) that display unique associations with DBR-CM items. The latter could address this concern by evaluating the specific or general improvements related to emphasizing the qualitative differences across items used to guide PD efforts (e.g., focusing on improving a discrete CM practice).

Conclusion

The use of evidence-based CM practices has been linked to a range of positive student outcomes, yet pre-service CM training for teachers has been largely neglected. In response, PD activities are increasingly using coaching, consultation, and PF, all of which rely on the valid and reliable assessment of teacher use of CM practices, to facilitate the development of CM skills. Present findings add to the growing evidence in support of the DBR-CM as a valid assessment of educator use of CM practices in varied contexts. For coaches and consultants in practice, preliminary DBR-CM validation efforts suggest that generated data are reliable and valid for use in support of data-driven PD activities. Although further accumulation of validity is warranted, the DCR-CM appears to be an emerging option for assessing educator behavior within tiered PD models (e.g., MTSES). When viewed within the broader DBR assessment methodology literature, the DBR-CM appears to be a highly usable, flexible assessment of teacher CM practices, making it a particularly desirable tool for coaching and PF-oriented PD activities.

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

This study was funded by an IES grant #R305A130143, Keith Herman, Principal Investigator and Wendy Reinke, co-Principal Investigator.

ORCID iD

Wesley A. Sims https://orcid.org/0000-0002-5431-0368

Supplemental Material

Supplemental material is available on the *Assessment for Effective Intervention* webpage with the online version of the article.

References

Alvero, A. M., Bucklin, B. R., & Austin, J. (2001). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). *Journal of Organizational Behavior Management*, *21*(1), 3–29. https://doi.org/10.1300/J075v21n01_02

Back, L. T., Polk, E., Keys, C. B., & McMahon, S. D. (2016). Classroom management, school staff relations, school climate, and academic achievement: Testing a model with urban high schools. *Learning Environments Research*, *19*(3), 397–410. https://doi.org/10.1007/s10984-016-9213-x

Begeny, J. C., & Martens, B. K. (2006). Assessing pre-service teachers' training in empirically validated behavioral instruction practices. *School Psychology Quarterly*, *21*(3), 262–285. https://doi.org/10.1521/scpq.2006.21.3.262

Bracken, S. S., & Fischel, J. E. (2006). Assessment of preschool classroom practices: Application of Q-sort methodology. *Early Childhood Research Quarterly*, *21*(4), 417–430.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education & Treatment of Children*, *34*(4), 575–591. https://doi.org/10.1353/etc.2011.0034

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of direct behavior rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention*, *34*(4), 201–213. https://doi.org/10.1177/1534508409340390

Christofferson, M., & Sullivan, A. L. (2015). Preservice teachers' classroom management training: A survey of self-reported training experiences, content coverage, and preparedness. *Psychology in the Schools*, *52*(3), 248–264. https://doi.org/10.1002/pits.21819

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement*, *12*(4), 425–434. https://doi.org/10.1177/014662168801200410

Cooper, J. T., Gage, N. A., Alter, P. J., LaPolla, S., MacSuga-Gage, A. S., & Scott, T. M. (2018). Educators' self-reported training, use, and perceived effectiveness of evidence-based classroom management practices. *Preventing School Failure: Alternative Education for Children and Youth*, *62*(1), 13–24. https://doi.org/10.1080/1045988X.2017.1298562

Flora, S. R. (2000). Praise's magic reinforcement ratio: Five to one gets the job done. *The Behavior Analyst Today*, *1*(4), 64–69. https://doi.org/10.1037/h0099898

Freeman, J., Simonsen, B. E., Briere, D. S., & MacSuga-Gage, A. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *The Journal of the Teacher Education Division of the Council for Exceptional Children*, *37*, 106–120. https://doi.org/10.1177/0888406413507002

Gage, N. A., MacSuga-Gage, A. S., & Crews, E. (2017). Increasing teachers' use of behavior-specific praise using a multitiered system for professional development. *Journal of Positive Behavior Interventions*, *19*(4), 239–251. https://doi.org/10.1177/1098300717693568

Gage, N. A., Scott, T., Hirn, R., & MacSuga-Gage, A. S. (2018). The relationship between teachers' implementation of classroom management practices and student behavior in elementary school. *Behavioral Disorders*, *43*(2), 302–315. https://doi.org/10.1177/0198742917714809

Grasley-Boy, N., Gage, N. A., & MacSuga-Gage, A. S. (2019). Multitiered support for classroom management professional development. *Beyond Behavior*, *28*(1), 5–12. https://doi.org/10.1177/1074295618798028

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Harrison, S. E., Riley-Tillman, T. C., & Chafouleas, S. M. (2014). Direct behavior rating: Considerations for rater accuracy. *Canadian Journal of School Psychology*, *29*(1), 3–20. https://doi.org/10.1177/0829573513515424

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Noell, G. H., Witt, J. C., LaFleur, L. H., Mortenson, B. P., Ranier, D. D., & LeVelle, J. (2000). Increasing intervention implementation in general education following consultation: A comparison of two follow-up strategies. *Journal of Applied Behavior Analysis*, *33*(3), 271–284. https://doi.org/10.1901/jaba.2000.33-271

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Upper elementary and secondary CLASS technical manual*. http://cdn2.hubspot.net/hubfs/336169/Technical_Manual.pdf

Reddy, L. A., Fabiano, G. A., & Dudek, C. M. (2013). Concurrent validity of the Classroom Strategies Scale for Elementary School—Observer form. *Journal of Psychoeducational Assessment*, *31*(3), 258–270. https://doi.org/10.1177/0734282912462829

Reinke, W. M., Herman, K. C., & Dong, N. (2018). The Incredible years teacher classroom management program: Outcomes from a group randomized trial. *Prevention Science*, *19*(8), 1043–1054. https://doi.org/10.1007/s11121-018-0932-3

Reinke, W. M., Sprick, R., & Knight, J. (2009). Coaching classroom management. In J. Knight (Ed.), *Coaching: Approaches & perspectives* (pp. 91–112). Corwin Press.

Reinke, W. M., Stormont, M., Herman, K. C., Wachsmuth, S., & Newcomer, L. (2015). The Brief Classroom Interaction Observation–Revised: An observation system to inform and increase teacher use of universal classroom management practices. *Journal of Positive Behavior Interventions*, *17*(3), 159–169. https://doi.org/10.1177/1098300715570640

Riley-Tillman, T. C., Chafouleas, S. M., Briesch, A. M., & Eckert, T. L. (2008). Daily behavior report cards and systematic direct observation: An investigation of the acceptability, reported training and use, and decision reliability among school psychologists. *Journal of Behavioral Education*, *17*(4), 313–327. https://doi.org/10.1007/s10864-008-9070-5

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A., & Glazer, A. D. (2008). Examining the agreement of direct behavior ratings and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions*, *10*(2), 136–143. https://doi.org/10.1177/1098300707312542

Riley-Tillman, T. C., Kalberer, S. M., & Chafouleas, S. M. (2005). Selecting the right tool for the job: A review of behavior monitoring tools used to assess student response to intervention. *The California School Psychologist*, *10*(1), 81–91. https://doi.org/10.1007/BF03340923

Schlientz, M. D., Riley-Tillman, T. C., Briesch, A. M., Walcott, C. M., & Chafouleas, S. M. (2009). The impact of training on the accuracy of Direct Behavior Ratings (DBR). *School Psychology Quarterly*, *24*(2), 73–83. https://doi.org/10.1037/a0016255

Shinn, M. R., & Shinn, M. M. (2002). *Administration and scoring of Reading Curriculum-Based Measurement (R-CBM) for use in general outcome measurement*. Edformation.

Simonsen, B., Freeman, J., Dooley, K., Maddock, E., Kern, L., & Myers, D. (2017). Effects of targeted professional development on teachers' specific praise rates. *Journal of Positive Behavior Interventions*, *19*(1), 37–47.

Simonsen, B., MacSuga-Gage, A. S., Briere, D. E., III, Freeman, J., Myers, D., Scott, T. M., & Sugai, G. (2014). Multitiered support framework for teachers' classroom-management practices: Overview and case study of building the triangle for teachers. *Journal of Positive Behavior Interventions*, *16*(3), 179–190. https://doi.org/10.1177/1098300713484062

Sims, W. A., King, K. R., Reinke, W. M., Herman, K., & Riley-Tillman, T. C. (2020). Development and preliminary validity evidence for the Direct Behavior Rating-Classroom Management (DBR-CM). *Journal of Educational and Psychological Consultation*, *312*(2), 215–245. https://doi.org/10.1080/10474412.2020.1732990

Sprick, R. S., Garrison, M., & Howard, L. M. (1998). *Champs: A proactive and positive approach to classroom management for grades K-9*. Sopris West.

Stichter, J. P., Lewis, T. J., Whittaker, T. A., Richter, M., Johnson, N. W., & Trussell, R. P. (2008). Assessing teacher use of opportunities to respond and effective classroom management strategies: Comparisons among high-and low-risk elementary schools. *Journal of Positive Behavior Interventions*, *112*(2), 68–81. https://doi.org/10.1177/1098300708326597

Tapp, J. (2002). *Multiple option observation system for experimental studies (MOOSES)* [Software]. http://mooses.vueinnovations.com/overview/mooses-overview

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving decision making in school psychology: Making a difference in the lives of students, not just a prediction about their lives. *School Psychology Review*, *47*(4), 385–395. https://doi.org/10.17105/SPR-2018-0042.V47-4

Wallace, T. L., Parr, A. K., & Correnti, R. J. (2020). Assessing teachers' classroom management competency: A case study of the Classroom Assessment Scoring System–Secondary. *Journal of Psychoeducational Assessment*, *38*(4), 475–492. https://doi.org/10.1177/0734282919863229

Wehby, J. H., Dodge, K. A., & Greenberg, M. (1993). *Classroom atmosphere rating scale. Unpublished technical manual*. Vanderbilt University.