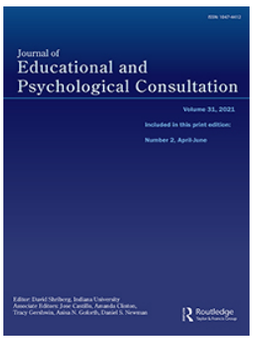

Routledge
Taylor & Francis Group

Journal of Educational and Psychological Consultation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hepc20

Development and Preliminary Validity Evidence for the Direct Behavior Rating-Classroom Management (DBR-CM)

Wesley A. Sims, Kathleen R. King, Wendy M. Reinke, Keith Herman & T. Chris Riley-Tillman

To cite this article: Wesley A. Sims, Kathleen R. King, Wendy M. Reinke, Keith Herman & T. Chris Riley-Tillman (2021) Development and Preliminary Validity Evidence for the Direct Behavior Rating-Classroom Management (DBR-CM), Journal of Educational and Psychological Consultation, 31:2, 215-245, DOI: 10.1080/10474412.2020.1732990

To link to this article: https://doi.org/10.1080/10474412.2020.1732990

JOURNAL OF EDUCATIONAL AND PSYCHOLOGICAL CONSULTATION
2021, VOL. 31, NO. 2, 215–245
https://doi.org/10.1080/10474412.2020.1732990

Routledge
Taylor & Francis Group

Development and Preliminary Validity Evidence for the Direct Behavior Rating-Classroom Management (DBR-CM)

Wesley A. Sims[a], Kathleen R. King[a], Wendy M. Reinke[b], Keith Herman[b], and T. Chris Riley-Tillman[b]

[a]University of California; [b]University of Missouri

ABSTRACT

Effective classroom management is of critical importance to the success of universal, Tier I supports and services. Unfortunately, teacher-reported deficits in classroom management training are well documented. In response, use of professional development activities such as consultation, coaching, and on-going performance feedback emphasizing skill-building have proliferated. The Direct Behavior Rating-Classroom Management (DBR-CM) was developed to facilitate screening and formative data collection to drive these activities. This study presents the background, development, and preliminary psychometric evidence for the DBR-CM. Specifically, this study examined inter-rater reliability and concurrent validity in support of the DBR-CM. Findings are promising with inter-rater reliability approaching or exceeding acceptable agreement levels and significant correlations noted between DBR-CM scores and concurrently completed measures of teacher classroom management behavior and perceived self-efficacy. Implications for use and future research are discussed, including further validation and refinement of the DBR-CM and its use within indirect, consultative service delivery.

ARTICLE HISTORY
Received 14 February 2019
Revised 30 January 2020
Accepted 17 February 2020

Effective delivery of universal, Tier I supports in schools is critical to promoting student academic, social, emotional, and behavioral success. In the integrated tiered service delivery approaches used in numerous educational settings today, the universal services provided at Tier I serve as the foundation for all service provision (Sugai & Horner, 2009). With the increase in use of tiered service delivery models, so too has scholarly attention examining factors supporting efficient and effective delivery of universal supports (Myers, Simonsen, & Sugai, 2011; Reinke et al., 2014). While the services and supports themselves (e.g., evidence-based curricula and programming) are important, concepts like implementation science, treatment integrity, and teacher accountability illustrate the importance of the processes and behavioral mechanisms used to deliver these services and supports. At Tier I, these processes and behavioral mechanisms may be best characterized as classroom management or a "holistic descriptor of

CONTACT Wesley A. Sims wesley.sims@ucr.edu University of California Riverside, 900 University Ave, 1207 Sproul Hall, Riverside, CA, 92506

teachers' actions in orchestrating supportive learning environments and building community" (Evertson & Harris, 1999, p. 60).

Despite the fundamental role it plays developing, shaping, and affecting change in student learning and behavior, research focusing on teacher behavior has received significantly less attention than school-based prevention, intervention, and remediation for student performance (Tillery, Varjas, Meyers, & Collins, 2010). This is unfortunate given the far- and wide-reaching impact of Tier I services and support for students. With this in mind, as the emphasis on teacher accountability and prevention within tiered service delivery models grows, so too has the transparency around supporting educator skill development and use. Increasingly, coaches, trainers, and consultants are being called upon to support use of evidence-based classroom management practices more explicitly. Consultation-based professional development activities seek to utilize the multiplicative effect of indirect service delivery when emphasizing teacher use of evidence-based classroom management practices likely to impact a significant number of children.

Since such efforts are often data-driven, the limited availability of assessment tools designed to assess educator classroom management behavior (Reddy, Fabiano, & Jimerson, 2013) may limit these consultative support efforts. High-quality data-based decision-making, including formative assessment decisions associated with coaching and performance feedback, is predicated on the use of reliable and valid assessments and resulting data. The validation process for new assessment tools is outlined by Kane (2013) and is described as an ongoing process of evidence accumulation relative to its intended uses and interpretations. The Direct Behavior Rating-Classroom Management (DBR-CM) is a novel application of the Direct Behavior Rating (DBR) assessment methodology to educator classroom management to address the limited availability of feasible, accessible, and defensible classroom management assessment options. This study outlines the initial interpretations and uses argument for the DBR-CM, including its use within consultation activities, and presents initial reliability and validity evidence to support proposed uses and interpretations.

Classroom management

Use of evidence-based classroom management practices has emerged as one prominent behavioral mechanism influencing the effectiveness of universal, Tier I services and supports (Simonsen, Fairbanks, Briesch, Myers, & Sugai, 2008). The importance of effective classroom management is based in the understanding that teaching academic and behavioral expectations, reinforcing expectations, anticipating and pre-correcting problem behavior before it happens, correcting inappropriate behavior respectfully, and establishing positive relationships are fundamental to promoting student success (Mitchell,

Hirn, & Lewis, 2017; Myers et al., 2011). Effective classroom management is linked to higher levels of student engagement and social competence (La Paro, Pianta, & Stuhlman, 2004; Pianta, La Paro, & Hamre, 2008), more time engaged in academic tasks, and increased levels of academic achievement (Brophy, 1988). Generally, well-managed classrooms are distinguished by educators' ability to monitor student attention and performance, establish behavioral expectations, and consistently implement rules and procedures that prevent problems from occurring (Reinke et al., 2014; Simonsen et al., 2008). In contrast, poorly managed classrooms are associated with higher levels of disruptive student behavior and lower levels of student on-task behavior and performance (Reinke, Lewis-Palmer, & Merrell, 2008). Beyond the apparent impact on student outcomes, poor classroom management practices have been linked to lower levels of teacher self-efficacy (Brouwers & Tomic, 2000) and higher levels of teacher burnout, stress, and turnover (Brouwers & Tomic, 2000; Friedman, 2013; Ozdemir, 2007). Unfortunately, while research has linked desirable student and teacher outcomes to effective classroom management practices, pre-service training programs have been slow to include explicit training in classroom management into curricula to any meaningful degree (Freeman, Simonsen, Briere, & MacSuga-Gage, 2014; Greenberg, Putman, & Walsh, 2014).

Consultation and classroom management

To address the deficiencies in pre-service classroom management training, districts and administrators have relied heavily on informational, in-service professional development activities that have had little impact on desired school improvement (Birman et al., 2007; Guskey, 2000; Guskey & Yoon, 2009). In an effort to shift professional development training objectives away from awareness-raising to behavior change, administrators and scholars are increasingly turning to coaching and on-going performance feedback embedded within a collaborative consultative framework (Mitchell et al., 2017; Reinke et al., 2008; Simonsen et al., 2017). Programs like the Classroom Check Up (Reinke et al., 2008), targeted professional development (Simonsen et al., 2017), Incredible Years (IY; Webster-Stratton, Reinke, Herman, & Newcomer, 2011), CHAMPS (Sprick, Garrison, & Howard, 1998), and PBIS plus (Hershfeldt, Pell, Sechrest, Pas, & Bradshaw, 2012) incorporate collaborative consultation practices into their specific coaching and performance feedback activities to support ongoing classroom management. Generally, collaborative consultation in schools is a process seeking to help classroom teachers in managing concerns by shifting their view of a concern from a "within-student" to a situational or environmental interaction perspective. A foundational component of a collaborative consultation framework is the perspective of equality within the consultee–consultant relationship, each with unique contributions to

potential solutions. The teacher is viewed as a skilled professional and the consultant brings expertise from a psychological perspective (Doveston & Keenaghan, 2010). In addition, this framework seeks to provide a supportive structure for problem-solving conversations that take place within the collaborative consultation process (Wagner, 2000).

Increasingly, all such conversations incorporate data. Inherent within coaching, performance feedback, and consultation approaches to professional development is the use of data to guide discussions around performance deficits and related improvement efforts (Reinke et al., 2008). Feasible, flexible, and psychometrically sound data facilitate the identification of support needed (i.e., screening), evaluating effectiveness of supports (i.e., formative assessment), the support process itself (i.e., performance feedback, coaching), and development and evaluation of supports themselves (i.e., empirical research cycle). Within coaching and performance feedback activities, a professional's qualified performance is considered explicitly, and goals and goal attainment are discussed. Whether initiated by an educator seeking support or identified through regular screening activities, using data to evaluate educator strengths and areas in need of improvement, and providing on-going support emphasizing skill development using this data has the potential to dramatically impact teacher performance and student outcomes (Jayaram, Moffit, & Scott, 2012; Reinke, Sprick, & Knight, 2009).

As the use of these professional development activities grows, so too does the need for technically adequate, efficient, and useful formative assessment tools with which to evaluate initial performance levels (i.e., screening) and growth (i.e., formative assessment; Simonsen et al., 2013). Unfortunately, despite increased work in this area, the number of accessible assessment instruments for educator classroom management practices remains limited (Reddy et al., 2013; Reinke, Stormont, Herman, Wachsmuth, & Newcomer, 2015).

Classroom management assessment

While representative of improvement in scholarly attention devoted to classroom management assessment, several identified classroom management assessments appear to have been developed for specific interventions or focus on a single or small number of behaviors of interest to specific studies (e.g., opportunities to respond, use of behavior-specific praise) rather than general classroom practices of interest to practitioners (Reddy et al., 2013). When selecting assessment instruments, targeted variables/behaviors, as well as the strengths, weaknesses, costs, availability, and benefits of all available tools must be carefully considered to determine which tool aligns best with the intended interpretations and uses and available resources (Riley-Tillman, Kalberer, & Chafouleas, 2005). Generally, four critical features of behavior assessment within a tiered service delivery approach include defensibility, flexibility, efficiency, and

repeatability (Chafouleas, Volpe, Gresham, & Cook, 2010). Additionally, Chafouleas and colleagues (2010) note feasibility as an important consideration for assessment selection. Although there is no explicit hierarchy for assessment selection considerations, it could be argued that feasibility, or considerations grounded in the resources required to access and use an assessment (Chafouleas et al., 2010), may hinder users from even considering the defensibility, flexibility, or repeatability of an assessment. Feasibility is influenced by several factors ranging from physical access to time, monetary costs, and person-power associated with using the assessment.

Traditional approaches to assessing teacher behavior in classrooms, when not inferred from student performance (Chetty, Friedman, & Rockoff, 2014), have frequently relied on either local rubrics, informal principal or coach observation notes, or teacher report via questionnaire or checklist (Bracken & Fischel, 2006; Reddy et al., 2013). Though well-intentioned, feasible, and economical, these approaches often lack standardization, objectivity, and sufficient validity evidence to support their use (Bracken & Fischel, 2006). As of late, in addition to these methods, educators and researchers have used systematic direct observation (SDO), rating scale, or hybrid SDO-rating scale assessment methodologies to assess classroom management behavior (Reinke et al., 2015). Primarily, for reasons related to psychometric defensibility and relatively low-level inference requirements, SDO has been noted as a particularly advantageous formative behavior assessment method in applied school settings (Riley-Tillman et al., 2005). Similarly, rating scales are often associated with higher levels of psychometric defeasibility (Riley-Tillman & Burns, 2010). Unlike SDO, however, rating scales often require a greater level of inference from raters, as they require users to aggregate subject behavior over numerous observations. As of late, scholars have attempted to access the strengths of these techniques by combining direct observation and rating scale assessment methodologies. This hybrid format organizes target behaviors into rating scale formatted rating systems to allow observers to rate observed behavior rather than track behavior along traditional dimensions (i.e., frequency, duration, rate, percentage).

Challenges facing available classroom management assessments

Although available literature describes the development, intended interpretations and uses, and some degree of validity evidence for existing classroom management measures, the protocols, forms, or software required to use these assessments can be inaccessible, as many require significant monetary or time expenditures for materials or training. For example, some available classroom management assessment tools require the purchase of both materials (e.g., manuals, protocols) and formal training delivered by certificated trainers. In other instances, some SDO options require investments in technology and software to obtain and analyze data. Once required technology

and software are acquired, additional training in both data collection and analysis is required. Given the well-documented funding challenges facing public schools, the associated costs suggest these may not be viable assessment tools for wide-spread use in school settings.

For some, rather than the time needed for training, the time needed to complete the assessment itself may render the assessment unusable. Some rating-scale formatted assessments are made of up a large number of items. Beyond time spent observing the subject, as the number of items increases, so too does the overall completion time. Again, this may be too time-consuming for regular use. Similarly, some assessments may require numerous exposures to subject performance to generate reliable data. Developers of one such assessment recommend six observation cycles to obtain reliable data (see Pianta et al., 2008), which may prove prohibitive. Ultimately, the limited number of accessible, feasible, defensible, and usable classroom management assessment tools may complicate selection of an assessment of classroom management behavior and hamper coach or consultant support efforts (Collier-Meek, Fallon, & Gould, 2018; Reddy, Dudek, & Shernoff, 2016; Reddy et al., 2013).

Direct Behavior Rating-Classroom Management (DBR-CM) development

Despite increased attention on classroom management, classroom management assessment, and consultation-based professional development activities, the availability of feasible, flexible, and defensible classroom management assessments remains limited (Reddy et al., 2013). The DBR-CM was developed to address this apparent shortage. Central to the use of any assessment is the confidence users have in the reliability and validity of the information it generates. The Interpretation/Use Argument approach to assessment validation outlines the non-linear processes for accumulating evidence to guide user confidence in assessments (Kane, 2013).

Interpretation/use argument validation approach

In an arguments-based approach to validation (Kane, 2013), assessment development begins by clearly outlining an argument around its intended interpretations and uses (interpretation/use argument, IUA; Kane, 2013). The intended interpretations and uses then guide the development, refinement, and validation process. Assessment developers accumulate evidence across four inference areas (i.e., Scoring, Generalization, Extrapolation, and Implications) that connect observation (i.e., data collection) to decision (see Cook, Brydges, Ginsburg, & Hatala, 2015; Kane, 2013, 1992). This contemporary approach to validation incorporates many of the traditional validation methods (i.e., Content, Criterion, and Construct validation). However, an arguments-based approach extends these methods by organizing this information around the inferences or decisions made using the information

obtained via a given assessment. Additionally, in this approach, an assessment is not considered valid based on a single piece of evidence (i.e., a study of construct validity). An assessment is considered valid when the accumulated evidence supporting the proposed interpretations and uses outweighs counterclaims (Cook et al., 2015). For example, as is the case here, early validation efforts compare obtained DBR-CM data to concurrent measures (e.g., SDO, CAS, and teacher self-efficacy) to accumulate validity evidence supporting extrapolation inferences (i.e., concurrent validity). This concurrent validity evidence becomes one component in the broader validation process which will include evidence supporting the additional claims or inferences outlined in the IUA for the DBR-CM (e.g., generalization inferences, implication inferences). Additional evidence addressing implication inferences (i.e., using DBR-CM data to identify professional development need) must be accumulated over time and multiple studies to reach a preponderance of evidence supporting the initial IUA (Cook et al., 2015).

DBR-CM IUA

Given the noted shortage in assessments of classroom management, the Direct Behavior Rating-Classroom Management (DBR-CM) was developed to serve as a feasible, defensible, and flexible assessment of classroom management. Specifically, the goal of DBR-CM development was to provide those individuals charged with identifying and supporting educator professional development needs (e.g., trainers, peer mentors, consultants, collaborators, and administrators) a feasible, defensible, flexible, efficient, repeatable, and usable screening and formative assessment of educator classroom management behavior. The DBR-CM seeks to provide screening and progress monitoring data indicative of educator use of evidence-based classroom management practices. Thusly, the validation process for the DBR-CM would endeavor to accumulate varied validity evidence such as predictive, concurrent, and convergent over several studies across varying participants and settings in support of this IUA. Additional empirical work would focus on implication inferences the examination of the utility and effects of embedding DBR-CM data in consultation practices. Specific practices would include screening for need, formative assessment during consultation process, and effects of inclusion of DBR-CM data on consultation outcomes (e.g., consultee functioning or student outcomes).

Direct behavior rating

To Support the IUA for the DBR-CM, developers relied heavily on a solid evidence base supporting the use of the direct behavior rating assessment methodology to assess behavior in educational settings (DBR; Chafouleas, 2011). The DBR assessment methodology is easily recognizable in daily behavior report cards, check in/check out, home-school notes, and good behavior

notes (Chafouleas, Riley-Tillman, & Sassu, 2006). In the last decade, scholars have added additional structure to DBRs to create DBR Single-Item Scales (DBR SIS; Chafouleas, Sanetti, Kilgus, & Maggin, 2012). From this work emerged a defensible (i.e., psychometrically sound) alternative behavior assessment methodology that combines the strengths of both rating scales and systematic direct observation (SDO) while incorporating principles of general outcome, formative assessment (see https://dbr.education.uconn.edu/library/publications/). Historically, the simple formatting and low inference completion requirements of DBRs allowed for efficient on-going collection of data (i.e., screening and formative assessment) that is easily organized for need identification, feedback, goal planning, reinforcement distribution, and other decision-making (Chafouleas, Riley-Tillman, & Christ, 2009b, 2009a) as well as use in monitoring one's own behavior (i.e., self-monitoring; Harrison, Riley-Tillman, & Chafouleas, 2014). The simple design of DBRs makes them an easy to understand and complete assessment that has proven to be a highly feasible method for the collection of behavioral data (Harrison et al., 2014). The DBR is also advantageous in that the design simplicity and ease of use require less training to use reliably in comparison to other formative assessment methods (e.g., systematic direct observation; Harrison et al., 2014; Riley-Tillman, Chafouleas, Briesch, & Eckert, 2008). The flexibility and efficiency of DBR SIS assessment format and methodology are well suited for the intended uses, screening (i.e., early identification of support need), and formative assessment (i.e., progress monitoring) of the DBR-CM as part of ongoing consultative activities.

Current study

This study serves as a necessary initial step in the accumulation of validity evidence in support of the DBR-CM for use by coaches, mentors, administrators, and consultants engaged in educator support and professional development activities. Use of assessments for educational decision-making is predicated on evidence establishing the reliability and validity of generated data. Accumulation of such data is guided by statements of the intended uses and interpretations of an assessment. Evidence relative to these claims and uses is then accumulated over time. To begin the accumulation of such evidence supporting the intended interpretations and uses of the DBR-CM, this study examines inter-rater reliability (i.e., generalization inferences) as well as concurrent validity (i.e., extrapolation inferences). This is to say, this study begins the accumulation of evidence to support the claim that the DBR-CM measures the construct of classroom management broadly, the individual components that make up this broad construct, and does so consistently across raters (i.e., reliably; variability in scores is related to subject performance). Since accumulation of such evidence begins an on-going process, this evidence is not meant to support use of the DBR-CM

relative to specific IUA goals, rather it should serve as preliminary support for use in low-stakes applications as well as justify additional scholarly efforts to validate specific claims relative to the IUA for the DBR-CM. Specific, study hypotheses include:

(1) In support of extrapolation inferences for the DBR-CM, significant positive correlations between the DBR-CM and concurrent SDO measures of educator classroom management behavior will be found, providing evidence of concurrent validity.
(2) Concurrent validity related to DBR-CM extrapolation inferences will be further evidenced through positive correlations between separate, concurrent behavior rating measures of educator classroom management behavior and perceived self-efficacy in classroom management.
(3) To support DBR-CM IUA generalization inferences, when using the DBR-CM, multiple observers will rate teacher classroom management behavior within acceptable levels of agreement (i.e., inter-rater reliability) using the DBR-CM.

Method

Participants

For the present study, data were collected across 107 classrooms in nine elementary schools in an urban Midwestern school district. Participating classroom educators included 107 regular education teachers in kindergarten through fourth grade. Demographic data collected indicated 94% were female. Reported racial makeup of participants was 79% Caucasian, 18% African American, 1% Asian, 1% Hispanic, and 1% other. Grade levels taught were 18% Kindergarten, 25% first grade, 22% second grade, 23% third grade, and 11% fourth grade. Reported age indicated 63% of participating teachers were between the ages of 20–40, with the remaining 37% falling above the age of 41. All participating teachers were certified by their state's department of education in elementary education (e.g., K-5.) Reported years of experience ranged from less than one to more than 30 years. Participation was voluntary and teachers completed an IRB approved consent process following initial recruitment.

Measures

Direct Behavior Ratings-Classroom Management

DBR-CM. Prior to the implementation of the current study, initial development of the DBR-CM occurred in four steps. First, a list of evidence-based classroom management practices was generated through a systematic review of available literature. Identified behaviors included but were not limited to

behavior-specific praise, general praise, opportunities to respond, use of pre-corrective statements (both academic and behavioral), instructional content delivery, reprimands, appropriate instructional pace, varied instructional methodologies, establishing rules and routines, student engagement, and using attention signals (Freeman et al., 2014; Pianta & Hamre, 2009; Reinke, Herman, & Stormont, 2013; Simonsen, Myers, & DeLuca, 2010; Solomon, Battistich, Kim, & Watson, 1996). Additionally, a common theme not captured by any distinct observable behavior emerged in the literature. A mutually positive, warm, and respectful relationship between the teacher and students as an essential feature of successful classrooms was repeatedly referenced in available classroom management literature, often emphasizing teacher efforts specifically to foster warmth and acceptance in classrooms (see Bracken & Fischel, 2006; Pianta & Hamre, 2009; Sprick et al., 1998).

Next, developers engaged in an informal sorting task where behaviors were grouped by similarity and commonality by behavior form and function independently before comparing, discussing, editing, and ultimately combining behavior groupings. Developers included a practicing school psychologist and doctoral-level student in school psychology and two Professors in School Psychology, one with expertise in classroom management and one with expertise in direct behavior ratings and assessment development and validation. Groupings were labeled in a functional yet creative manner and operational definitions were developed (see Table 1). Resulting groupings (i.e., DBR-CM items) included Praise, Communication, Engagement, Enthusiasm, and Rapport. To facilitate observer attention to instances of desired or positive behavior and to align with item scoring (e.g., higher scores equate to more instances of positive classroom management behaviors), operational definitions were worded with emphasis on positive or proactive behaviors or actions.

Third, operational definitions were condensed to short, succinct priming definitions and placed into a DBR SIS format. This format includes a 0 to 10, Likert-style rating scale for each item and its associated priming definition. The DBR-CM was formatted onto a single sheet of paper with items and demographic information scoring on the front and completion directions and operational definitions on the back. This form is available for viewing on the Open Science Framework page for this project (https://osf.io/5kaz9/?view_only=1f70f5eec5f842e783c5139e3ee37573, link anonymized for peer review).

Finally, a conceptual model using identified groups and themes noted in literature was developed to illustrate DBR-CM development and broadly outlined goals. Ideally, assessment using the DBR-CM will drive professional development activities, through consultation, and in turn increase successful students at Tier I and beyond. This framework is also available for viewing on the Open Science Framework page for this project (see anonymized for peer review project page at https://osf.io/5kaz9/?view_only=1f70f5eec5f842e783c5139e3ee37573). This along with

the DBR-CM form were reviewed by developers and a panel of reviewers to identify potential revisions. The review panel included two classroom management researchers, two assessment researchers, one classroom educator, one practicing school psychologist, and two advanced doctoral-level graduate students in school psychology. Noteworthy revisions included the addition of an evaluation of classroom structure and minor semantic changes to the operational definitions of DBR-CM items. This process ended with review panel members using the form to score video examples of teacher behavior. Members reported the DBR-CM directions were clear, the form was easy to use, and inter-rater agreement high.

The development process resulted in an assessment that utilizes 0 to 10, Likert-style rating scale to rate each of five broad classroom management items. These individual item scores can be summed to compute an overall classroom management score. The DBR-CM assesses five core elements of classroom management: praise, communication, engagement, rapport, and enthusiasm. Each item encompasses several discrete behaviors that are similar to one another in form or function (see Table 1).

Direct observation of teacher behavior

BCIO-R. The BCIO-R (Reinke et al., 2015) is a behavioral observation coding system completed by external classroom observers. In this study, BCIO-R observations were completed using MOOSES software (Tapp, 2002) and handheld computer devices. MOOSES software allows for temporally sequenced frequency data for targeted behaviors, which adds additional strength to inter-rater reliability information by adding a temporal sequence component to analyses (see Analysis section for additional information). Teacher target behaviors included in this study were: educator use of behavior-specific praise, general praise, pre-corrections, opportunities to respond, explicit reprimands, harsh reprimands, time teaching, and time not teaching during each observation. Frequency data gathered using the BCIO-R can be converted to rate (i.e., number of occurrences per minute). BCIO-R data may also be used to calculate an indicator of overall classroom management, the "Positive Implementation" variable. This variable is used to represent overall classroom management and serves as an indicator of the quality of teacher implementation of effective classroom management practices (Reinke et al., 2014). Mean percent agreement on the BCIO-R has ranged from 88% (0–100%) to 90% (79–100%; Reinke et al., 2014). Several BCIO-R teacher behaviors have been found to be significantly related with one another, with observed correlations ranging from ($r = 0.19$, $p < .05$) to ($r = 0.36$, $p < .01$; Reinke et al., 2014).

Global observation scale of classroom management

Classroom atmosphere scale. The Classroom Atmosphere Scale (CAS; Wehby, Dodge, & Greenberg, 1993) is used to assess the quality of an

Table 1. Direct Behavior Rating-Classroom Management items and associated operational definitions.

Item	Item Definition
Praise	Praise is the use of positive statements or actions, including distribution of tangible reinforcers, in response to the behavior and performance of students in the classroom. In the classroom, Praise looks like: Educator uses more behavior-specific praise than general praise, uses praise contingent on expected behavior, provides three (3) or more praise statements for every reprimand, reprimands are few and when used are not harsh, educator is more positive than negative when interacting with students, provides praise at desirable rates using non-verbal interactions such as gestures, tangibles, or physical contact, and maintains an overall tone that is positive and not negative or sarcastic.
Communication	Communication refers to the clear communication of goals and expectations of an instructional period. In the classroom, Communication looks like: Educator provides clear academic and behavioral expectations to the students, explicitly states or posts instructional objectives and offers opportunity for clarifying questions, clearly presents behavioral expectations verbally and/or visually, uses an attention signal to gain attention of all students, and utilizes transition procedures that appear to be known and followed by majority of students (as evidenced by efficient classroom transitions).
Engagement	Engagement is 90% or more of students engaged 80% of the time during instruction and/or classroom activities and students are provided and respond to questions posed to the group and individual students occur frequently. In the classroom, engagement looks like 90% of students are clearly academically engaged at all times; level of observable disruptions in the classroom is minimal: teacher provides four (4) or more opportunities for students to respond per minute during instruction; and teacher asks many different students in the classroom at least one question during instruction.
Enthusiasm	Enthusiasm is the delivery of instructional content in a meaningful, memorable, and/or engaging manner. In the classroom, enthusiasm is evident when the educator's tone and pace of instruction are positive and upbeat, instructional content is supplemented with or related to a familiar life applications, topics, or activities, and instruction incorporates alternative activities (e.g. students as teachers, group work, pair and share, current event, etc.).
Rapport	Rapport is the quality of the student-educator relationship, especially that of mutual trust, emotional affinity, acceptance, and positivity. In the classroom, Rapport looks like: The general feel in the classroom is mutually warm and accepting; the educator uses children's names frequently; interactions between the educator and students are visibly positive; the educator answers clarification questions posed by students; and the educator appears to feel comfortable, positive, and genuine in his/her interactions with students.

instructional environment. CAS items are scored on a 5-point Likert scale, ranging from 1 to 5, with higher ratings indicating more positive classroom atmosphere and lower scores indicating more negative atmosphere. Classroom atmosphere items rated on the CAS include levels of student compliance during structured times, compliance during transitions, adherence to rules, cooperation, interest and engagement, on-task behavior, and the degree to which the environment was supportive of student behavior. Items are scored in the same direction and are used to calculate an overall score representing overall classroom atmosphere. Standard alpha coefficients of .94 to .95 establish good internal consistency for the CAS and moderate interrater reliability, with a reported interclass correlation coefficient of .44 (n = 115; Barber, Maggin, & Wehby, 2009; Wehby et al., 1993).

Teaching efficacy

The Ohio state teacher efficacy scale (OSTES). The OSTES (Tschannen-Moran & Hoy, 2001) measures teacher perceptions of self-efficacy. Teachers completed eight items comprising the classroom management subscale. Teachers responded to each item by indicating their perceptions of self-efficacy on a 0 to 9 scale, with lower scores indicating lower perceived self-efficacy. Cronbach's alpha for the subscale ranged from .95 to .96 (Tschannen-Moran & Hoy, 2001).

Procedure

Data used in this study were collected as part of a large-scale efficacy trial for the Incredible Years teacher classroom management training program (Webster-Stratton et al., 2011). This grant-funded study was managed by a research and resource center attached to a large research-intensive university in the Midwest. Data used in this study were collected during a single observation period (i.e., one observation per participant) in the spring, at least one-year post-treatment initiation. Participant intervention status was blinded from observers throughout the study. Following development, addition of the DBR-CM to study measures was approved by the institutional review boards (IRB) the primary researchers' institution as well as the participating school district prior to its use. Approximately 15 graduate research assistants, principal investigators, and university-based research center staff conducted data collection observations.

Data collection training. To facilitate data collection, observers participated in training for all assessments included in the study, though the depth and breadth of this training varied by measure. BCIO-R training consisted of approximately 1 hour of didactic training and 30 minutes of group practice with video clips. For the BCIO-R, observers completed training and reliability checks using videos and practice sessions in live classrooms over a 2-week period to meet a minimum criterion level of 85% reliability with a master coder. All observers for the data collected for this study reached and maintained this criterion level. Prior to data collection for this study, operational definitions of CAS items and DBR-CM items were reviewed and discussed with data collectors. Regarding actual ratings of DBR-CM items, consistent with DBR SIS, observers were advised to use an anchor system. Observers were told that a score of 10 should represent the presence of the operational definition. As behavior is observed that is inconsistent with the operational definition or as components of the operational definition go unobserved, raters would remove points from the ideal score of 10.

Data collection. Classroom observations were completed during a 20-minute instructional period (e.g., reading or math) for each participating

teacher. During this instructional period, direct BCIO-R observation data were collected by one observer. Following each BCIO-R observation, the observer completed either the DBR-CM or CAS. A second observer completed the remaining measure (i.e., either the DBR-CM or CAS). This is to say, in one condition Observer 1 completed the BCIO-R and then the DBR-CM and Observer 2 completed the CAS. In a second condition, Observer 1 would complete the BCIO-R and the CAS, and Observer 2 would complete the DBR-CM. Participating teachers completed the OSTES along with any additional demographic data questionnaires separately from the observation period, though in conjunction with the data collection period.

Inter-observer agreement (IOA) data were collected for 34% of the primary outcome measures used in the study (i.e., IOA data were collected for 34% of BCIO-R and DBR-CM). The mean percentage agreement for the BCIO-R was 91%, ranging from 67% to 93%. MOOSES utilizes second-by-second comparison of raters to determine reliability; an overall reliability of 80% is considered acceptable, thus 91% is considered highly reliable (Tapp, 2002).

Analytic plan and anticipated findings

First, descriptive statistics for each measure were calculated. Results indicated that a minimal amount of data appeared to be missing. Since missing data were limited to only a few data points, replacement procedures were not used.

Next, concurrent validity was examined using bivariate correlations between the DBR-CM and the BCIO-R, CAS, and OSTES. Significant positive correlations were expected between DBR-CM item scores and concurrent measure items or behaviors, with the exception of reprimands. DBR-CM items are all worded positively; thus, only significant negative correlations were expected for items of negative behaviors (i.e., reprimands) or negatively worded items. Similar findings are also anticipated for DBR-CM total scores and total or global scores for concurrent measures.

Finally, inter-rater reliability statistics were computed for the DBR-CM using two methods, percent agreement and Intra-class Correlation Coefficient (ICC). Additionally, for DBR-CM data, percent agreement was calculated using two methods, exact agreement and (±) 1-point. For exact agreement, scores for two raters had to be exactly the same to be considered in agreement (e.g., rater 1 DBR-CM Praise = 6 and rater 2 DBR-CM = 6 represented agreement). For the (±) 1-point method, DBR-CM data ratings were considered to be in agreement if they varied by no more than 1-point (e.g., rater 1 DBR-CM Praise = 6 and rater 2 DBR-CM = 7 represents agreement). This second method is consistent with recommendations provided by DBR SIS developers. Acceptable levels of inter-rater reliability were anticipated for the DBR-CM.

Results

Descriptive data for DBR-CM, BCIO-R, and CAS variables are presented in Table 2. Generally speaking, scores on the DBR-CM and CAS fell around the mid-point of their respective scales. Ratings of teacher behaviors were not exceptionally high or low, indicating observers recorded acceptable and homogenous levels of positive classroom management practices. DBR-CM item ratings resulted in mean scores ranging from 5.9 to 7.3 (SD = 2.14 to 2.5), with median scores ranging from 7 to 8. These scores appeared to cluster at the upper end of the middle rating options (i.e., approximately 4 to 7 on a 0 to 10 scale). Similar findings were noted in CAS ratings. Mean CAS item ratings ranged from 2.92 to 3.67 (SD = .91 to 1.13), while mean ratings ranged from 3 to 4 on the 0 to 5 scale. Rates of BCIO-R variables ranged from .01 to 1.6 (SD = .42 to 1.97). These results indicate that target behaviors (e.g., praise, reprimands, precorrections) were typically observed at a rate of less than one per minute, with the exception of opportunities to respond (OTR, M = 1.6, SD = 1.13). Finally, overall, educators responses yielded a mean efficacy score of 7.70 (SD = .92; median = 7.75). Generally, these data indicate limited variability in observer or rater perceptions of

Table 2. Descriptive statistics for study variables.

	N Valid	*N* Missing	Mean	Median	Std. Deviation
Direct Behavior Rating-Classroom Management – Praise	107	3	5.9	7.00	2.41
Direct Behavior Rating-Classroom Management – Communication	107	3	6.2	7.00	2.14
Direct Behavior Rating-Classroom Management – Engagement	106	4	7.3	8.00	2.20
Direct Behavior Rating-Classroom Management – Enthusiasm	107	3	6.1	7.00	2.50
Direct Behavior Rating-Classroom Management – Rapport	107	3	6.7	7.00	2.32
Direct Behavior Rating-Classroom Management – Total	106	4	32	34.0	10.35
Brief Classroom Interaction Observation-Revised – Positive implementation	105	2	60	61.54	23.75
Brief Classroom Interaction Observation-Revised – Rate of overall praise	105	2	.72	.55	.60
Brief Classroom Interaction Observation-Revised – Rate of precorrection	105	2	.01	.00	.032
Brief Classroom Interaction Observation-Revised – Rate of opportunities to respond	105	2	1.6	1.00	1.97
Brief Classroom Interaction Observation-Revised – Rate of overall reprimands	105	2	.47	.40	.42
Classroom Atmosphere Scale – Compliance	107	0	3.63	4.00	.98
Classroom Atmosphere Scale – Rules	107	0	3.58	4.00	.92
Classroom Atmosphere Scale – Cooperation	106	1	3.67	4.00	.91
Classroom Atmosphere Scale – Interest	107	0	3.50	3.00	.99
Classroom Atmosphere Scale – Focused	107	0	3.59	4.00	1.05
Classroom Atmosphere Scale – Individual differences	107	0	2.92	3.00	1.13
Classroom Atmosphere Scale – Supportive	107	0	3.45	3.00	1.00
Ohio State Teacher Efficacy Scale – Efficacy	107	0	7.70	7.75	.92

classroom management practices use or efficacy. Mean and median scores appeared similar and standard deviations, relative to scaling for respective assessments, indicating perceptions of limited variability in ratings overall.

To address Hypotheses 1 and 2, bivariate correlations between DBR-CM items and other classroom management variables were calculated and are presented in Tables 3–6. First, to address Hypothesis 1, which anticipated DBR-CM ratings would be positively related to SDO variables, bivariate correlational analyses compared DBR-CM items and BCIO-R variables. Results indicated overall DBR-CM scores are significantly positively correlated with the BCIO-R Positive Implementation variable (r = .53, p = .01; see Table 3). Furthermore, the BCIO-R Positive Implementation variable was significantly correlated with all individual DBR-CM items (see Table 4). Not surprisingly, the largest correlations were noted between DBR-CM Praise (r = .56, p = .01) and Rapport (r = .52, p = .01) items and the BCIO-R Positive Implementation variable.

Similarly, significant positive correlations were noted between the BCIO-R Rate of praise variable and DBR-CM Praise (r = .57, p = .01), Communication (r = .40, p = .01), Enthusiasm (r = .33, p = .01), and Rapport (r = .24, p = .05) items. No significant correlations were evident between DBR-CM items and the BCIO-R rate of precorrect variable. A significant positive correlation was noted between DBR-CM Enthusiasm ratings and the BCIO-R Opportunities to respond variable (r = .21, p = .05). As anticipated, several significant correlations were evident between DBR-CM items and the BCIO-R rate of reprimands variable. All DBR-CM items were found to be significantly negatively correlated with this variable except for DBR-CM Praise (r = .17, p = .01).

Table 3. Intercorrelations among overall scores on classroom atmosphere scale, brief classroom interaction observation – Revised, and Direct Behavior Rating-Classroom Management measures.

	Direct Behavior Rating-Classroom Management – Total	Classroom Atmosphere Scale – Total	Brief Classroom Interaction Observation-Revised – Positive implementation	Ohio State Teacher Efficacy Scale – Total
Direct Behavior Rating-Classroom Management – Total	1			
Classroom Atmosphere Scale – Total	.81**	1		
Brief Classroom Interaction Observation-Revised – Positive implementation	.53**	.52**	1	
Ohio State Teacher Efficacy Scale – Total	.25**	.37**	.30**	1

** Correlation is significant at the 0.01 level.

Table 4. Intercorrelations among overall scores on brief classroom interaction observation-revised and Direct Behavior Rating-Classroom Management measures.

	1	2	3	4	5	6	7	8	9	10	11
1. Direct Behavior Rating-Classroom Management – Total	1										
2. Direct Behavior Rating-Classroom Management – Praise	.84**	1									
3. Direct Behavior Rating-Classroom Management – Communication	.89**	.75**	1								
4. Direct Behavior Rating-Classroom Management – Engagement	.87**	.56**	.72**	1							
5. Direct Behavior Rating-Classroom Management – Enthusiasm	.94**	.70**	.82**	.79**	1						
6. Direct Behavior Rating-Classroom Management – Rapport	.92**	.69**	.72**	.80**	.84**	1					
7. Brief Classroom Interaction Observation-Revised – Positive Implementation	.53**	.56**	.39**	.49**	.38**	.52**	1				
8. Brief Classroom Interaction Observation-Revised – Rate of praise	.39**	.57**	.40**	.19	.33**	.24*	.46**	1			
9. Brief Classroom Interaction Observation-Revised – Rate of precorrect	.04	−.03	.05	.09	.07	.01	.15	.23*	1		
10. Brief Classroom Interaction Observation-Revised – Rate of OTR	.16	.14	.12	.15	.21*	.08	.07	.04	.05	1	
11. Brief Classroom Interaction Observation-Revised – Rate all reprimands	−.33**	−.17	−.22*	−.43**	−.20*	−.44**	−.59**	.12	−.03	.02	−.02

*Correlation is significant at the 0.05 level.
**Correlation is significant at the 0.01 level.

Table 5. Intercorrelations among overall scores on classroom atmosphere scale and Direct Behavior Rating-Classroom Management measures.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Direct Behavior Rating-Classroom Management – Total	1												
2. Direct Behavior Rating-Classroom Management – Praise	.84**	1											
3. Direct Behavior Rating-Classroom Management – Communication	.89**	.75**	1										
4. Direct Behavior Rating-Classroom Management – Engagement	.87**	.56**	.72**	1									
5. Direct Behavior Rating-Classroom Management – Enthusiasm	.94**	.70**	.82**	.79**	1								
6. Direct Behavior Rating-Classroom Management – Rapport	.92**	.69**	.72**	.80**	.84**	1							
7. Classroom Atmosphere Scale – Compliance	.59**	.41**	.44**	.69**	.49**	.62**	1						
8. Classroom Atmosphere Scale – Rules	.68**	.48**	.52**	.75**	.60**	.70**	.85**	1					
9. Classroom Atmosphere Scale – Cooperation	.65**	.45**	.52**	.69**	.55**	.66**	.75**	.75**	1				
10. Classroom Atmosphere Scale – Interest	.77**	.55**	.65**	.72**	.77**	.72**	.65**	.69**	.67**	1			
11. Classroom Atmosphere Scale – Focused	.71**	.47**	.59**	.78**	.59**	.71**	.83**	.83**	.76**	.69**	1		
12. Classroom Atmosphere Scale – Individual difference	.59**	.49**	.52**	.48**	.56**	.56**	.40**	.43**	.45**	.51**	.46**	1	
13. Classroom Atmosphere Scale – Supportive	.77**	.64**	.59**	.67**	.72**	.76**	.67**	.68**	.66**	.72**	.67**	.57**	1

**Correlation is significant at the 0.01 level.

Table 6. Mean percentage agreement, interclass correlation coefficient values, and range of reliability for direct behavior rating-classroom management.

Variable	Mean Percentage Agreement	Percentage Agreement (± 1)	Interclass Correlation Coefficient	Interclass Correlation Coefficient Lower Limit	Interclass Correlation Coefficient Upper Limit
Direct Behavior Rating-Classroom Management – Praise	.69	.69*	.83	.66	.91
Direct Behavior Rating-Classroom Management – Communication	.75	.81*	.67	.34	.83
Direct Behavior Rating-Classroom Management – Engagement	.69	.69*	.75	.51	.87
Direct Behavior Rating-Classroom Management – Enthusiasm	.78	.78*	.83	.67	.91
Direct Behavior Rating-Classroom Management – Rapport	.67	.67*	.84	.68	.92
Direct Behavior Rating-Classroom Management – Total	.70	-	.79	.60	.89

*Only calculated for individual DBR-CM items.

Next, Hypothesis 2 anticipated positive correlations between the DBR-CM and concurrent behavior rating measures of educator classroom management or reported self-efficacy. DBR-CM scores were found to be significantly positively correlated with CAS scores and the OSTES efficacy score and are reported in Tables 3 and 5. The DBR-CM and CAS total scores appeared to be significantly positively correlated (r = 0.81, $p < .01$; see Table 3). Similarly, a significant positive correlation was also found between the DBR-CM total score and the OSTES total score (r = 0.25, $p < .01$; see Table 3). Significant positive correlations of varying strengths were noted between all individual DBR-CM and CAS items. Correlation values ranged from .41 to .78 ($p = .01$).

Hypothesis 3 stated that acceptable levels of inter-rater agreement (i.e., >.70) would be evident in DBR-CM ratings. To address this hypothesis inter-rater reliability statistics were computed and are reported in Table 6. Inter-rater reliability statistics approached or exceeded the desired .70 level (Cronbach, 1951) for DBR-CM items. Mean percentage agreement ranged from .67 to .78 for exact agreement between DBR-CM ratings. Mean percentage agreement values exceeded the .70 level for DBR-CM Communication, Enthusiasm, and Total scores. Values approached but did not meet or exceed the .70 threshold for DBR-CM Praise, Engagement, and Rapport items. Mean percentage agreement values ranged from .69 to .81 when calculated using the ± 1 method. Again, values for DBR-CM Communication and Enthusiasm items exceeded the .70 threshold, whereas DBR-CM Praise and Engagement items did not. The DBR-CM Total score was excluded from this calculation as no precedent for evaluating its reliability has been established previously. ICC values ranged from .67 to .84. ICC values for all items except DBR-CM Communication exceeded the .70 level.

Discussion

This study serves as the introduction to the DBR-CM, an assessment of educator use of evidence-based classroom management practices. This study outlines the development process, the Interpretation Use Argument that guides the validation process, and preliminary validity evidence for the DBR-CM. Overall, findings appear promising, as this initial step focuses on addressing generalization and extrapolation inferences for the IUA. This is to say, this study sought to establish that DBR-CM ratings measure educator classroom management behavior consistently (i.e., inter-rater reliability) and as intended (i.e., concurrent validity).

First, descriptive statistics for each measure indicated relatively homogenous classroom management practices. Results suggest raters perceived educator classroom management practices as generally positive and lacking significant variability overall (i.e., few instances of either exemplary high or significantly poor low performance) as measured by these assessments. The

tight clustering of scores around the middle-most rating options, or slightly higher, and muted standard deviations relative to the range of rating options used, suggests observers did not view educator classroom management practices as significantly deficient, generally. This seems contrary to expectations given the consistent and persistent reports of deficits in pre-service training in classroom management (Freeman et al., 2014; Simonsen et al., 2013). The reasonable anticipation that general educator difficulty in their use of classroom management practices (i.e., scores lower than 5 on DBR-CM items, lower than 3 on CAS items) would be reflected in ratings was not confirmed.

Relatedly, teacher self-reports of classroom management efficacy indicate more positive perceptions of functioning in this area. Overall, responses place beliefs of classroom management efficacy at the upper (i.e., positive) end of the rating scale, with little apparent variability in these scores. Based on the challenges apparently experienced and reported by teachers noted previously, one could again reasonably anticipate findings that cluster at the lower end of the performance scale (i.e., more scores that indicate problematic functioning).

In contrast to these measures, descriptive statistics for SDO data present a less positive view of classroom management practices than other measures used in the study. For instance, overall, findings suggest that educators provide 1.5 praise statements (i.e., general and specific combined) for every reprimand. This falls well below the 4:1 ratio recommended by prior researcher (Myers et al., 2011). Furthermore, overall rates of pre-corrective statements indicate this evidence-based proactive prevention strategy is rarely used by classroom educators (0.01 per minute). Greater variability is noted in the overall rate of opportunities to respond used by educators in this study. Unfortunately, on average, educators provided opportunities to respond at a rate (1.6 per minute) well below the 4 to 6 per minute recommended by the effective instruction literature (CEC, 1987).

The apparent contradiction evident in these data may be explained as a function of a variety factors. First, it is possible that optimal thresholds for frequencies and rates of evidence-based classroom management practices may be significantly disconnected from actual practice in the applied setting. For instance, the 4:1 praise to reprimand ratio espoused by Myers et al. (2011) may be unattainable for most teachers or in some environments. Additional explanation for the generally positive perspective of teacher classroom management on Likert formatted assessments and apparently less favorable findings from SDO assessments may be related to the inferences required by these assessments. Advocates of SDO methodology would likely argue that this format is more objective, an observer either sees a target behavior or does not. In contrast, Likert-style rating formats allow more subjectivity in their ratings. This allows for biases, both positive and negative,

to influence observer ratings. It is possible observers for this study had slightly more positive or favorable perceptions of teachers generally or these participants specifically and these favorable perceptions were reflected in their ratings. Similarly, the apparent generally favorable ratings on self-reported measures of self-efficacy may be related to a general desired to not be looked at negatively in response to assessment results, inflation of one's actual functioning.

Hypotheses guiding this study focused on expected agreement of the DBR-CM items, including the total score, with concurrent measures of educator use of classroom management practices. These hypotheses were guided by extrapolation inferences (i.e., scores reflect actual performance) outlined by the IUA approach for assessment development and validation. Overall, preliminary findings of the DBR-CM appear promising. Correlational analyses found evidence of concurrent validity when DBR-CM Total scores are compared to concurrent measures of classroom management broadly (i.e., BCIO-R Positive Implementation, CAS Total score, and OSTES Self-efficacy) . Correlations between these broad measures of classroom management were positive and significant suggesting as a general or broad concept, classroom management is assessed by the DBR-CM similarly to other broad or general measures of classroom management.

Similarly, further correlational analyses indicated evidence of concurrent validity for the individual DBR-CM items. Significant correlations of varying strength were noted between DBR-CM items and all individual CAS items. As hypothesized, correlations were in expected directions, as both CAS and DBR-CM items are positive behaviors or worded positively. Significant positive correlations were noted between all DBR-CM and CAS items. These results may indicate these measures do not identify discreet classroom management behaviors. The significant positive correlation values found between all CAS and DBR-CM variables could be an indication that these measures of classroom management may not distinguish between the broad concept of classroom management and its subcomponents. This is to say that each individual item on of these measures may not contribute uniquely to the assessment of classroom management. Given the small number of items comprising these measures, alternative analyses (i.e., exploratory factor analysis and principal components analysis) typically used to answer such questions are difficult. While the pervasive agreement is problematic in that it did not discriminate between individual's behaviors and practices, it is again indicative of concurrent validity for the DBR-CM generally, suggesting additional work to address this question as it relates to the extrapolation inferences for the DBR-CM IUA is warranted.

Significant correlations in expected directions were also noted between individual DBR-CM items and individual BCIO-R variables. Unlike CAS findings, not all items were significantly positive correlated. As hypothesized

negative correlations were also evident for several DBR-CM variables and the BCIO-R reprimand variables. This was as expected given the emphasis on positively stating operational definitions and scoring on the DBR-CM. Few significant correlations were evident between DBR-CM items and the BCIO-R opportunities to respond and pre-correction variables. This may be a function of including both process (i.e., teacher behavior) and outcome (i.e., student behavior) components in the operational definition for engagement and the overlap between these two DBR-CM items. This suggests the DBR-CM items, primer definitions, and full operational definitions will likely need revision as work with the DBR-CM continues.

Study findings suggest inter-rater reliability approached or exceeded acceptable levels for DBR-CM items. For each calculation, inter-rater reliability levels neared or met desired reliability levels in spite of a relatively brief training consisting of exposure to and discussion of the operational definitions of DBR-CM items. This may be indicative of the minimal inferencing need and high feasibility often touted by DBR-based assessments, though these claims have yet to be specifically evaluated for the DBR-CM. Given previous work with similar assessments, reliability levels would likely meet or exceed desired levels with additional training and reliability checks (see Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009).

Overall, the nature of these results is promising given the noted levels of reliability, moderate to strong correlations noted between concurrently conducted classroom management assessments, and teacher-reported self-efficacy. While these data are preliminary in nature, correlations between broad scores and individual items of the DBR-CM and concurrent measures in the appropriate directions support continued work in the accumulation of validity evidence for this measure.

Limitations

Though promising, these preliminary findings are not without limitations. First, the cursory nature of the training observers received in the use of the DBR-CM is a noted limitation. While the ease of accurate, reliable use may be advantageous for practitioners, in an empirical study such as this, care should be taken to ensure data collectors are thoroughly prepared to reliably collect data. Though such training was unavailable at the time of this study, it is reasonable to expect a more thorough training and reliability check process will improve DBR-CM levels of interrater agreement. In future studies, thorough DBR-CM training and reliability checks should occur prior to data collection.

An additional limitation of this study is noted in the completion of the BCIO-R, DBR-CM, and CAS by the same observer in some instances immediately following completion of the observation period. It is possible these

measures acted as a confounding or priming influence on the subsequently completed measures. In applied research, investigators must weigh the benefits of limiting the confounding influence of observers in natural settings with the detriments posed by restricting the number of unique observers collecting data concurrently. In this study, it was determined that the benefits of additional observers to complete assessments independently did not outweigh the potential disadvantages (i.e., fewer concurrent observers were deemed more advantageous). Future studies should, if possible, attempt to include completion of assessment tools independently (i.e., by independent observers) to limit potential priming of confounding effects.

The assessments used to examine concurrent validity in this study may also be a limitation of this study. As noted, the cost and limited availability of many alternative classroom management assessments necessitated the use of the BCIO-R, CAS, and OSTES in this study. Ideally, concurrent validation would include additional and varied assessment tools, in particular the CLASS (Pianta et al., 2008) and CSS (Reddy, Fabiano, & Dudek, 2013). Generally, consistent with the ongoing nature of an IUA approach to validation, use of additional concurrent measures across multiple studies supports the accumulation of validity evidence (i.e., more concurrent measures would likely lead to the accumulation of more evidence of concurrent validity). Furthermore, including measures that have greater empirical and face validity, such as the CLASS and CSS, serves to lend additional credence to the resulting validity evidence.

Finally, the inclusion of process and outcome behaviors in the operational definition of the DBR Engagement variable may have impacted findings. The outcome behavior (i.e., apparent student engagement) should likely be removed from this assessment completely or from the operational definition of DBR-CM Engagement item. The small but significant correlation between the DBR-CM Enthusiasm item and the BCIO-R "opportunities to respond" item supports this argument. This suggests teacher behavior currently subsumed in the "Engagement" item (e.g., opportunities to respond) may be better grouped as part of the "Enthusiasm" item.

Implications for research and practice

Given the reported deficits in pre- and in-service professional development, in conjunction with a growing emphasis on prevention within a tiered service delivery approach, coaches, consultants, and trainers are increasingly being called upon to support improvement in educator classroom management practice. Use of performance feedback and coaching incorporating screening and formative assessment data imbedded within a collaborative consultation framework attempts to shift professional development goals away from awareness-raising to behavior change (Mitchell et al., 2017; Reinke et al., 2008; Simonsen et al., 2017). The DBR-CM was developed to address the

limited availability of feasible classroom management assessments. Kane (2013) describes the validation of any new assessment as the accumulation of evidence to support its proposed uses and interpretations. This study begins the validation process for the DBR-CM, through the accumulation of scoring, generalization, and extrapolation evidence through the examination of reliability and concurrent validity (Cook et al., 2015; Kane, 2013).

These preliminary findings represent an initial step in the accumulation of validity evidence to support the proposed interpretations and uses of the DBR-CM. Future work with the DBR-CM should continue the accumulation of validity evidence to support use within a multi-tiered system of educator support (MTSES). Specifically, the primary proposed interpretations and uses of the DBR-CM are the efficient, feasible, and defensible collection of screening and progress monitoring data for use in professional development activities, particularly coaching and performance feedback. This work should be replicated and extended by incorporating additional grade levels, samples, and concurrent validation assessments.

Additional work should focus on the accumulation of evidence supporting scoring inferences through further exploration of inter-rater reliability. Specifically, the impact of a more rigorous training and reliability check process should be explored. To support extrapolation inferences, work around the diagnostic accuracy of the DBR-CM should be conducted. Future work should explore the diagnostic accuracy of the DBR-CM total and individual items by examining sensitivity and specificity, negative-, and positive-predicative power relative to classroom management classification instruments (e.g., the CLASS, CSS). Finally, the ultimate goal of any classroom management assessment is to identify educator practices that impact student outcomes. Future work should examine extrapolations and implication inferences by exploring the relationships between DBR-CM scores and a variety of student outcomes (e.g., academic performance, social-emotional and behavioral functioning).

In addition to the psychometric properties of the DBR-CM, future studies investigate how consultants and coaches could utilize the data generated by the DBR-CM to guide or inform their consultation practices with educators. Future work should explore feasibility, perceived utility, and effects of consultative practices incorporating DBR-CM data. A highly touted aspect of the DBR assessment methodology is the ease with which it can be used (Riley-Tillman, Methe, & Weegar, 2009). It is often reported to be a more user-friendly behavior assessment option, particularly for practitioner use in screening and formative assessment. Future work should explicitly explore feasibility of DBR-CM use in applied settings as a component of a collaborative consultation process, in which consultants support teacher training and professional development to improve use of evidence-based classroom management practices. Furthermore, beyond feasibility, both

perceived (i.e., social validity) and actual effects (i.e., efficacy) of DBR-CM use within consultative practices such as coaching and performance feedback should be explored. In addition to defensibility (i.e., psychometric and validity evidence), determining the value of DBR-CM data to consultant and consultee users is a critical question that should be explored. Ultimately, the goal of any consultative activity is to effect positive change in a collaboratively identified area difficulty. Future work should examine the effects of consultative practices incorporating DBR-CM data on stated goals for the consultative relationship, likely classroom management or other related outcomes (e.g., student behavior or achievement).

Conclusion

As the use of multi-tiered service delivery models expands, so to do efforts to increase the efficiency and effectiveness of universal, Tier I services and supports. Classroom management, or educator efforts to oversee the activities of the classroom, has emerged as an impactful factor in the effectiveness of Tier I service delivery. Unfortunately, training in evidence-based classroom management practices appears neglected in teacher pre-service training. To address this challenge facing classroom educators, professional development activities are increasingly utilizing performance feedback, coaching, and consultation to facilitate skill development in this and other areas. These contemporary approaches to professional development require the collection of reliable and valid screening and formative data, in a feasible manner. This study served as a preliminary examination of the newly developed DBR-CM. The application of the DBR SIS assessment methodology to classroom management is timely and relevant given the increasing use of tiered service delivery models in schools, renewed examination of factors impacting service delivery, and increased use of coaching and performance feedback to support teachers in schools. Though additional examination is needed, the results of this study suggest the DBR-CM may be a promising data collection option for those who support educator use of evidence-based classroom management practices.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Institute of Education Sciences [R305A100342].

References

Barber, B., Maggin, D., & Wehby, J. (2009). *Improving the reliability and validity of classroom atmosphere assessment: the classroom atmosphere rating scale–revised*. In Presentation at AERA Annual Convention.

Birman, B., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., … O'Day, J. (2007). State and local implementation of the no child left behind act [Product page]. Retrieved from https://www.rand.org/pubs/reprints/RP1283.html

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. doi:10.1016/j.ecresq.2006.09.006

Brophy, J. (1988). Educating teachers about managing classrooms and students. *Teaching and Teacher Education*, *4*(1), 1–18. doi:10.1016/0742-051X(88)90020-0

Brouwers, A., & Tomic, W. (2000). A longitudinal study of teacher burnout and perceived self-efficacy in classroom management. *Teaching and Teacher Education*, *16*(2), 239–253. doi:10.1016/S0742-051X(99)00057-8

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. doi:10.1353/etc.2011.0034

Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2009a). Direct Behavior Rating (DBR). *Assessment for Effective Intervention*, *34*(4), 195–200. doi:10.1177/1534508409340391

Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2009b). Direct Behavior Rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention*. Retrieved from http://psycnet.apa.org/psycinfo/2009-12305-001

Chafouleas, S. M., Riley-Tillman, T. C., & Sassu, K. A. (2006). Acceptability and reported use of daily behavior report cards among teachers. *Journal of Positive Behavior Interventions*, *8*(3), 174–182. doi:10.1177/10983007060080030601

Chafouleas, S. M., Sanetti, L. M., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change using Direct Behavior Rating single-item scales. *Exceptional Children*, *78*(4), 491–505. doi:10.1177/001440291207800406

Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010). School-based behavioral assessment within problem-solving models: current status and future directions. *School Psychology Review*, *39*(3), 343-349.

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. doi:10.1257/aer.104.9.2593

Collier-Meek, M. A., Fallon, L. M., & Gould, K. (2018). How are treatment integrity data assessed? Reviewing the performance feedback literature. *School Psychology Quarterly*, *33*, 517–526. doi:10.1037/spq0000239

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. doi:10.1111/medu.12678

Council for Exceptional Children. (CEC). (1987). *Academy for effective instruction: working with mildly handicapped students*. Reston, VA: Author.

Cronbach, L. J. (1951). *Coefficient alpha and the internal structure of tests. Psychometrika, 16*(3), 297–334.

Doveston, M., & Keenaghan, M. (2010). Teachers and educational psychologists working together: What can we learn? *Support for Learning*, *25*(3), 131–137. doi:10.1111/j.1467-9604.2010.01451.x

Evertson, C. M., & Harris, A. H. (1999). Support for managing learning-centered classrooms: The classroom organization and management program. *Beyond behaviorism:Changing the classroom management paradigm*, 59–74.

Freeman, J., Simonsen, B., Briere, D. E., & MacSuga-Gage, A. S. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *Teacher Education and Special Education*, *37*(2), 106–120. doi:10.1177/0888406413507002

Friedman, I. A. (2013). Classroom management and teacher stress and burnout. In Evertson, C. M., & Weinstein, C. S. (Eds.). *Handbook of classroom management* (pp. 935–954). New York, NY: Routledge.

Greenberg, J., Putman, H., & Walsh, K. (2014). *Training our future teachers: Classroom management. National council on teacher quality*. Washington, DC: National Council on Teacher Quality. Retrieved February 2018, from https://files.eric.ed.gov/fulltext/ED556312.pdf

Guskey, T. R. (2000). *Evaluating professional development*. Retrieved from http://books.google.com/books?hl=en&id=CklqX4zgDtgC&oi=fnd&pg=PR9&dq=ineffective+professional+development+in+schools&ots=gS1vDK0Ahu&sig=W7zrAZT0w4zdvl0otgQrACVpcd8

Guskey, T. R., & Yoon, K. S. (2009). What works in professional development? *Phi Delta Kappan*, *90*(7), 495–500. doi:10.1177/003172170909000709

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. doi:10.1177/0829573513515424

Hershfeldt, P. A., Pell, K., Sechrest, R., Pas, E. T., & Bradshaw, C. P. (2012). Lessons learned coaching teachers in behavior management: The PBIS plus coaching model. *Journal of Educational and Psychological Consultation*, *22*(4), 280–299. doi:10.1080/10474412.2012.731293

Jayaram, K., Moffit, A., & Scott, D. (2012). Breaking the habit of ineffective professional development for teachers. *McKinsey.[Links]*. Retrieved from http://www.centralsquarefoundation.org/wp-content/uploads/2016/02/Breaking-the-habit-of-ineffective-professional-development-for-teachers.pdf

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. doi:10.1037/0033-2909.112.3.527

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, *104*, 409–426. doi:10.1086/499760

Mitchell, B. S., Hirn, R. G., & Lewis, T. J. (2017). Enhancing effective classroom management in schools: Structures for changing teacher behavior. *Teacher Education and Special Education*, *40*(2), 140–153. doi:10.1177/0888406417700961

Myers, D. M., Simonsen, B., & Sugai, G. (2011). Increasing teachers' use of praise with a response-to-intervention approach. *Education and Treatment of Children*, *34*(1), 35–59. doi:10.1353/etc.2011.0004

Ozdemir, Y. (2007). The role of classroom management efficacy in predicting teacher burnout. *International Journal of Social Sciences*, *2*(4), 257–263.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119. doi:10.3102/0013189X09332374

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom assessment scoring system. *Baltimore: Paul H. Brookes*. Retrieved from http://www.elcmdm.org/QualityCounts/SPMMay23.pdf

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

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. doi:10.1177/0734282912462829

Reddy, L. A., Fabiano, G. A., & Jimerson, S. R. (2013). Assessment of general education teachers' Tier 1 classroom practices: Contemporary science, practice, and policy. *School Psychology Quarterly*, *28*(4), 273. doi:10.1037/spq0000047

Reinke, W. M., Herman, K. C., & Stormont, M. (2013). Classroom-level positive behavior supports in schools implementing SW-PBIS identifying areas for enhancement. *Journal of Positive Behavior Interventions*, *15*(1), 39–50. doi:10.1177/1098300712459079

Reinke, W. M., Lewis-Palmer, T., & Merrell, K. (2008). The classroom check-up: A classwide teacher consultation model for increasing praise and decreasing disruptive behavior. *School Psychology Review*, *37*(3), 315.

Reinke, W. M., Sprick, R., & Knight, J. (2009). Coaching classroom management. In Knight, J. (Ed.) *Coaching: Approaches & Perspectives* (pp. 91–112). Thousand Oaks, CA: 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. doi:10.1177/1098300715570640

Reinke, W. M., Stormont, M., Herman, K. C., Wang, Z., Newcomer, L., & King, K. (2014). Use of coaching and behavior support planning for students with disruptive behavior within a universal classroom management program. *Journal of Emotional and Behavioral Disorders*, *22*(2), 74–82. doi:10.1177/1063426613519820.

Riley-Tillman, T. C., & Burns, M. K. (2010). *Evaluating educational interventions: Single-case design for measuring response to intervention*. New York, NY, US: Guilford press.

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. doi:10.1007/s10864-008-9070-5

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. doi:10.1007/BF03340923

Riley-Tillman, T. C., Methe, S. A., & Weegar, K. (2009). Examining the use of Direct Behavior Rating on formative assessment of class-wide engagement A case study. *Assessment for Effective Intervention*, *34*(4), 224–230. doi:10.1177/1534508409333879

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. doi:10.1037/a0016255

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, *31*(3), 351–380. doi:10.1353/etc.0.0007

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. doi:10.1177/1098300716637192

Simonsen, B., MacSuga-Gage, A. S., Briere, D. E., Freeman, J., Myers, D., Scott, T. M., & Sugai, G. (2013). 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. doi:10.1177/1098300713484062.

Simonsen, B., Myers, D., & DeLuca, C. (2010). Teaching teachers to use prompts, opportunities to respond, and specific praise. *Teacher Education and Special Education: the Journal of the Teacher Education Division of the Council for Exceptional Children*, *33*(4), 300–318. doi:10.1177/0888406409359905

Solomon, D., Battistich, V., Kim, D., & Watson, M. (1996). Teacher practices associated with students' sense of the classroom as a community. *Social Psychology of Education*, *1*(3), 235–267. doi:10.1007/BF02339892

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

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, *17*(4), 223–237. doi:10.1080/09362830903235375

Tapp, J. (2002). *Multiple option observation system for experimental studies (MOOSES) [Software]. Unpublised Technical Manual, Vanderbilt University, Nashville, Tennessee*. Retreived from http://mooses.vueinnovations.com/sites/default/files/public/d7/moosesmanual_0.pdf.

Tillery, A. D., Varjas, K., Meyers, J., & Collins, A. S. (2010). General education teachers' perceptions of behavior management and intervention strategies. *Journal of Positive Behavior Interventions*, *12*(2), 86–102. doi:10.1177/1098300708330879

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805. doi:10.1016/S0742-051X(01)00036-1

Wagner, P. (2000). Consultation: Developing a comprehensive approach to service delivery. *Educational Psychology in Practice*, *16*(1), 9–18. doi:10.1080/026673600115229

Webster-Stratton, C., Reinke, W. M., Herman, K. C., & Newcomer, L. L. (2011). The incredible years teacher classroom management training: The methods and principles that support fidelity of training delivery. *School Psychology Review*, *40*(4), 509.

Wehby, J. H., Dodge, K. A., & Greenberg, M. (1993). Classroom atmosphere rating scale. *Unpublished Technical Manual*, Vanderbilt University, Nashville, Tennesee.

Notes on contributors

Wesley A. Sims, Ph.D., NCSP, is an Assistant Professor in the School Psychology program at the University of California, Riverside's Graduate School of Education. His research activities include improving educator service delivery practices within tiered service delivery systems, implementation science in educational settings, and assessment of educator classroom management behavior. Dr. Sims began his career as a practicing School Psychologist where he served a variety of schools and populations, and garnered extensive experience facilitating support services activities within tiered service delivery models.

Kathleen R. King, Ph.D., NCSP, is a Visiting Assistant Professor in the School Psychology Program at UCR. Her research focuses on the development and validation of behavioral observation and screening measures within a broader problem-solving/MTSS framework. In addition to her work targeting student behavior, Dr. King is involved in efforts to evaluate and improve teacher practices in the applied setting.

Wendy M. Reinke, Ph.D., is a Professor in the Educational, School, & Counseling Psychology department at the University of Missouri and co-director of the Missouri Prevention Science Institute. Her primary research interests in training and supporting school personnel to

deliver effective practices with a particular interest in personnel use of evidence-based social behavioral and emotional interventions. She has been the PI or Co-PI on over $38 million in federal research grants, many of which have focused on supporting teachers use of effective classroom management practices. She has co-authored seven books, including Motivational Interviewing for Effective Classroom Management: The Classroom Check-up, eleven chapters and over 100 peer-reviewed publications related to prevention of social emotional and behavior problems.

Keith Herman, Ph.D., is a Curator's Distinguished Professor in Department of Education, School, & Counseling Psychology at the University of Missouri. He is the Co-Founder and Co-Director of the Missouri Prevention Science Institute. He has an extensive grant and publication record including over 120 peer-reviewed publications in the areas of prevention and early intervention of child emotional and behavior disturbances and culturally-sensitive education interventions.

T. Chris Riley-Tillman, Ph.D., is a Professor and Associate Provost at the University of Missouri. He is one of the co-developers of Direct Behavior Ratings as well as a recognized authority in evidence-based practice in schools and the application of experimental design and analysis in applied educational settings. Related to these interests, Dr. Riley-Tillman has participated in leadership roles on seven federal grants and is a Senior Advisor for the National Center on Intensive Intervention. He is also the creator and lead developer of the Evidence Based Intervention Network.