Journal of School Psychology 68 (2018) 53–72

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Methods matter: A multi-trait multi-method analysis of student behavior☆

Faith G. Miller[a,*], Austin H. Johnson[b], Huihui Yu[c], Sandra M. Chafouleas[c], D. Betsy McCoach[c], T. Chris Riley-Tillman[d], Gregory A. Fabiano[e], Megan E. Welsh[f]

[a] University of Minnesota, United States
[b] University of California – Riverside, United States
[c] University of Connecticut, United States
[d] University of Missouri, United States
[e] University at Buffalo, United States
[f] University of California – Davis, United States

ARTICLE INFO

Action Editor: Andy Garbacz

Keywords:
Assessment
Behavior
Direct Behavior Rating
MTMM
Systematic direct observation
Validity
Rating scales

ABSTRACT

Reliable and valid data form the foundation for evidence-based practices, yet surprisingly few studies on school-based behavioral assessments have been conducted which implemented one of the most fundamental approaches to construct validation, the multitrait-multimethod matrix (MTMM). To this end, the current study examined the reliability and validity of data derived from three commonly utilized school-based behavioral assessment methods: Direct Behavior Rating – Single Item Scales, systematic direct observations, and behavior rating scales on three common constructs of interest: academically engaged, disruptive, and respectful behavior. Further, this study included data from different sources including student self-report, teacher report, and external observers. A total of 831 students in grades 3–8 and 129 teachers served as participants. Data were analyzed using bivariate correlations of the MTMM, as well as single and multi-level structural equation modeling. Results suggested the presence of strong methods effects for all the assessment methods utilized, as well as significant relations between constructs of interest. Implications for practice and future research are discussed.

1. Introduction

Students who develop emotional and behavioral disorders experience some of the poorest academic, social, and behavioral outcomes of any disability group (Bradley, Doolittle, & Bartolotta, 2008). As such, remediation of such difficulties is essential. The use of a prevention-focused approach that begins with using psychometrically sound assessment methods to support early identification, intervention, and problem solving for these students, followed by the delivery of evidence-based interventions and systems that actively support and enhance implementation of interventions, have the potential to substantially improve these students' trajectories. Reliable and valid assessment data form the foundation of such evidence-based practices; decisions and inferences are only as good as the available data. To this end, clear understanding of the strengths and limitations of our assessment data is essential. The

☆ Preparation of this article was supported by funding provided by the Institute of Education Sciences, U.S. Department of Education (R324A110017: PI, Chafouleas). Opinions expressed herein do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred.

* Corresponding author at: University of Minnesota, Department of Educational Psychology, 250 Education Sciences Building, Minneapolis, MN 55455, United States.
E-mail address: fgmiller@umn.edu (F.G. Miller).

https://doi.org/10.1016/j.jsp.2018.01.002
Received 14 October 2016; Received in revised form 7 December 2017; Accepted 24 January 2018

purpose of this study was to examine the reliability and validity of school-based behavioral assessment data targeting three constructs (i.e., academic engagement, disruptive behavior, and respectful behavior) which may serve as critical indicators of student success in schools.

1.1. Remediating behavioral problems in educational settings

Given the substantial negative academic, behavioral, and social-emotional outcomes associated with problem behavior, researchers and practitioners have focused upon the identification of both prevention and intervention techniques targeted towards behavioral problems. The prevention framework that has been integrated into systems of service delivery across disciplines (Institute of Medicine, 2012) focuses broadly on avoiding problems through universal promotive strategies and early intervention with individuals who begin to exhibit risk for problems. This framework is reflected in formalized multi-tiered systems of support (MTSS) implemented by schools to promote student behavioral success through prevention and intervention, such as School-Wide Positive Behavior Interventions and Support (SWPBIS; Sugai & Horner, 2002) and Safe and Civil Schools (Sprick, Booher, & Garrison, 2009). Although the identification and implementation of effective practices to support the behavioral success of all students is critical, the decision-making process involved in determining student need, the identification of an intervention best suited to an individual student's presenting problem, and the ongoing monitoring of student responses to these interventions depend on the use of assessment data which demonstrate evidence for their validity and reliability in that specific decision-making context.

Behavioral assessment data are essential to understanding whether prevention and intervention services are warranted and effective (Deno, 1986). Accordingly, within MTSS, methods for behavioral data collection vary based upon the question of concern. At the whole-school level, data sources such as office discipline referrals (ODRs) and standardized screeners (e.g., Walker, Cheney, Stage, & Blum, 2005) may provide actionable information on levels of need and groups or individuals in need of additional support. At the small-group or individual-student level, where interventions are being implemented in order to remediate behavioral challenges before they become clinically significant or to treat specific behavioral concerns, data sources are often focused on progress monitoring for the purpose of evaluating student response to intervention (Kazdin, 2005). Relevant assessment methods include those that rely on a teacher or another professional's observation of discrete student behavior, such as systematic direct observation (Suen & Ary, 1989) and Direct Behavior Rating (Chafouleas, Sanetti, Kilgus, & Maggin, 2012), as well as rating scales which assess behavior from a larger retrospective period of time (e.g., Gresham et al., 2010). Given the central role that such data play in multi-tiered decision-making, the psychometric quality of behavioral assessment data is a critical consideration for researchers and practitioners who seek to improve behavioral outcomes for students.

1.2. Reliable and valid behavioral assessment data

The appropriateness of traditional psychometric standards and analyses as applied to behavioral assessment, and indeed the defining features of behavioral assessment itself, has been fraught with debate (Silva, 1993). Many researchers placed direct observation, accompanying the development of Applied Behavior Analysis and its related methodologies, in contrast with traditional methods of personality and cognitive assessment which had heretofore defined the psychological and educational assessment landscape. As described by Silva (1993), the introduction of direct observation as a cornerstone of assessment in a behavioral paradigm was accompanied by the proliferation of measures which were criticized as having little to no evidence for the quality of the resulting data.

As part of these concerns, extensive interest arose in determining precisely how to evaluate the quality of behavioral assessment tools, with significantly differing views presented across multiple decades of research. As Silva (1993) described, “It is rare when we feel we have ‘incontrovertible’ measures of anything, and in a philosophical sense incontrovertible standards simply do not exist. No matter how fine-grained the measurement, we always only approximate reality” (p. 498). With that history acknowledged, it is worth revisiting two foundational concepts within psychometrics, reliability and validity, in order to provide context for both and define our position for each within behavioral assessment intended for use in decision-making within multi-tiered frameworks.

1.2.1. Reliability and validity

Reliability, broadly defined as a metric of consistency in scores along some dimension of interest, can be gauged using a variety of methods (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Given the distinct assumptions made by different assessment methods, the particular approach used to estimate a reliability coefficient should vary by assessment format. For example, when scores are repeatedly collected over time and aggregated, one might use an intraclass correlation to estimate the consistency of the collection of scores. This reflects an assumption that this aggregated score should be derived from individual observations which reflect systematic change or consistency in an object of measurement, rather than unsystematic change or error. In contrast, interrater reliability estimates such as the agreement index or the kappa coefficient might be used for observational assessment among multiple raters, when consistency within observations across raters is the topic of interest. Finally, when methods involve both observation and multiple timepoints, multiple reliability metrics can be used.

Perhaps the most important notion with regard to reliability, however, is the key role that it plays in validation. Validity, or the extent to which it is appropriate to use a test in a particular way, is at the heart of most psychometric arguments, and scores cannot be valid if they are not reliable because unreliable scores will not yield a consistent signal about students' behavioral status. However, reliability is only the first step. To establish that it is appropriate to use an assessment for screening or progress monitoring,

researchers must also collect foundational evidence that ascertains the extent to which a construct is adequately captured by a measure and, according to modern validity theorists, should also examine intended and unintended consequences of assessment (Messick, 1995) or delineate all conditions that must hold to support a particular use and gather evidence with regard to those conditions (Kane, 1992).

The application of validity to the behavioral realm has largely focused on the notion of construct validity instead of on modern validation theory (as discussed in Silva, 1993). As such, it focuses on the extent to which assessment scores reflect behavioral adjustment or maladjustment instead of how the scores are used. Construct validity can be gauged in a variety of ways including examining test content, the extent to which the cognitive processes used to generate scores are consistent with testing expectations, relations of test scores with other variables, and the interrelationships between items on the same measure (AERA, APA, & NCME, 2014).

Perhaps the most rigorous approach to construct validation, however, was described by Cronbach and Meehl (1955) when they proposed validation through the use of nomological nets. This involves explicating the network of relationships that should exist between: (a) different observable indicators, (b) constructs and the observable indicators, and (c) different constructs, and then empirically testing the extent to which these relationships hold. Multi-trait multi-method matrices (MTMM), discussed next, were first proposed by Campbell and Fiske (1959) as an elegant approach to operationalizing nomological nets, and continue to be discussed in measurement texts as a rigorous validation technique (e.g., McCoach, Gable, & Madura, 2013), although they are rarely implemented in behavioral assessment.

1.3. MTMM

Theoretical assumptions about the inter-relationships between constructs and the methods used to measure them can be tested using MTMM. By explicitly outlining the expected relationships between various constructs and the measures used to gauge them, researchers can evaluate the extent to which these hypothesized relationships, whether strong or weak, are satisfied by data. The MTMM organizes these relationships and hypotheses into a matrix of correlation coefficients, the statistical derivation of which has expanded beyond simple bivariate correlations into structural equation modeling through confirmatory factor analytic methods (e.g., Joreskog, 1971; Widaman, 1985). The use of confirmatory factor analyses in MTMM analyses also permits the partitioning of measurement error from the targeted sources of variance, strengthening the resulting conclusions (Hofling, Schermelleh-Engel, & Moosbrugger, 2009).

Matrices in MTMM analyses provide convergent and discriminant evidence, as encouraged by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014). In practical terms, convergent evidence refers to the degree to which data from two measures targeting the same construct agree (or correlate), whereas discriminant evidence is concerned with the degree to which data from two measures targeting distinct constructs disagree (or do not correlate). The values present within an MTMM can therefore be categorized into convergent validity coefficients, discriminant validity coefficients, and reliability coefficients. These coefficients are organized into distinct areas within the matrix. First, validity diagonals consist of validity coefficients depicting the correlations between the same trait measured using distinct methods, also known as monotrait-heteromethod values. Discriminant validity coefficients are organized into heterotrait-monomethod and heterotrait-heteromethod triangles, which depict correlations among measures of different constructs using the same method, and measures of different constructs using different methods, respectively. Finally, reliability coefficients are displayed where monotrait-monomethod coefficients would be placed (similar to the diagonal of a standard correlation matrix) in order to depict the relative reliability of the data derived from each individual measure used in the analysis (Campbell & Fiske, 1959).

In the years between the proposal of MTMM and revision of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), scholars debated the appropriateness of using MTMM to validate behavior assessment. Cone (1976) permitted that MTMM are valuable to behavior research if they focus on the examination of relationships between behaviors and limit inferences about constructs, but adopted a much more conservative perspective in later years – even questioning the value and relevance of traditional psychometric procedures for behavioral assessment. However, as described previously, Silva's (1993) argument that constructs are inextricable to behavior assessment and indeed to science itself, makes a forceful claim for the utility of MTMM. Silva adopted a more progressive view, and argued that MTMM may provide a window into both “the definition of the construct and the nomological network into which it is inserted” (p. 112). Thus, whether approached from a conservative or progressive view of psychometrics and its application to behavioral assessment, authors on both sides have suggested that MTMM analyses may provide critical information on the quality of behavioral assessments. This need is particularly salient when applied to school-based settings, given the significant implications of behavior problems for students and schools. Unfortunately, to date, few investigations have used MTMM analyses to examine relationships among constructs and measures of student behavior.

1.4. Behavioral constructs, assessment methods, and raters

1.4.1. Behavioral constructs

As previously discussed, the notion of constructs underlies validity in both traditional and behavior assessment. Given the substantial educational, professional, and social implications associated with child and adolescent behavior problems (e.g., Masten et al., 2005; Patterson, Forgatch, Yoerger, & Stoolmiller, 1998), the identification and assessment of specific constructs which are critical indicators of behavioral problems and success is critical. In one initial framework, the keystone intervention model, four behaviors have been posited as integral to the development of and transfer to other skills: compliance, social skills, on-task skills, and

communication skills (Ducharme & Shecter, 2011). Such indicators have been broadly referred to in the literature as general outcome measures, or GOMs. Some consensus has been reached regarding GOMs in academic domains (e.g., Hosp, Hosp, & Howell, 2007), leading to guiding principles for the identification of constructs which act as GOMs in behaviors; these should be "(a) relevant to all students in classrooms and (b) relevant influences on potential for (or barriers to) student academic success" (Chafouleas, 2011, p. 582).

Three behaviors which may provide a critical window into general student behavior functioning include academically engaged, disruptive, and respectful behaviors. Indeed, academic engagement has been suggested as a core behavioral indicator of academic success through its role as an "academic enabler" (DiPerna & Elliott, 2002); this construct may be critical given the negative educational outcomes observed for students with behavior problems. In addition, disruptive and respectful behavioral constructs may work to exemplify the two fundamental types of disruptive behavior disorders (i.e., antisocial and defiant/disrespectful behavior patterns) posited by Gresham (2015), and compliance (which is operationalized in a manner parallel to "respect" in this study) is identified as one of the four critical keystone behaviors described by Ducharme and Shecter (2011). Within an MTMM framework, each of these constructs are conceptualized as distinct traits that can be evaluated independent of the methods used to measure them.

1.4.1.1. Academically engaged behavior. Generally defined as active or passive participation in an academic activity (Shapiro, 2004), academic engagement exists within a set of behaviors also referred to as academic enablers within the educational and psychological literature (DiPerna & Elliott, 2002). A substantial amount of literature exists to support the predictive and concurrent validity of academically engaged behavior as it correlates with academic achievement (for a review, see Greenwood, Horton, & Utley, 2002). Thus, in the search for behaviorally-based and academically-relevant GOMs, it appears that academic engagement is a key construct to consider due to its relevance to some of the most salient school outcomes. Like compliance/respect, on-task behaviors are another of the four critical behaviors in the keystone framework (Ducharme & Shecter, 2011). School-based assessment of academic engagement may include direct observation measures such as the Behavioral Observation of Students in Schools (BOSS; Shapiro, 2011) and subscales of rating scales such as the Social Skills Improvement System's Motivation to Learn (SSIS; Gresham & Elliott, 2008).

1.4.1.2. Disruptive behavior. Behavior problems are often conceptualized as those which interfere with students and their educational and social systems (Bambara, 2005), and students who engage in higher-than-average levels of problem behaviors experience significant deleterious outcomes during childhood, adolescence, and into adulthood. Thus, disruptive behavior, consisting of those behaviors which disrupt classroom functioning, is another key construct to examine in relation to school-based behavior assessment. As a construct, disruptive behavior has been defined in numerous ways: delimited along psychopathological diagnoses such as Oppositional Defiant Disorder and Conduct Disorder (American Psychiatric Association, 2013), conceptually disaggregated into oppositional, aggressive, and hyperactive behaviors (e.g., Stormshak, Bierman, McMahon, Lengua, & Conduct Problems Prevention Research Group, 2000), or more broadly described as externalizing behaviors distributed into subtypes such as delinquent and aggressive behaviors, as in the Child Behavior Checklist (e.g., Greenbaum & Dedrick, 1998). Given the numerous ways disruptive behavior has been operationalized and conceptualized within the literature, this study defined disruptive behavior as those behaviors which interfere with typical classroom function.

This specific focus on behaviors which disrupt typical classroom functioning may be of particular interest to teachers and educators as they seek to provide a supportive and functional learning environment, and these behaviors may be related to teacher burnout (Fernet, Guay, Senécal, & Austin, 2012). The wide range of topographies of behavior which may disrupt typical classroom functioning can subsequently be measured in a variety of ways, including rating scales like the Behavioral Assessment System for Children, Third Edition (BASC-3; Reynolds & Kamphaus, 2015).

1.4.1.3. Respectful behavior. Respectful behavior, defined in this study as compliant and polite behavior in response to adult direction and/or interactions with peers and adults, is a critical component of many educators' perceptions of classroom functioning (e.g., Bridgeland, Bruce, & Hariharan, 2013). However, use of the term respectful for an overarching construct is not without difficulty. For instance, behaviors which disrupt classroom functioning may simultaneously be perceived as disrespectful in some definitions of this construct, and indeed, some researchers and measures use the word "respect" alongside determinations of disruption (e.g., Fernet et al., 2012; Gresham, 2015). Within respect (or its negative converse, disrespect), defiance, noncompliance, and refusal may potentially comprise some components within researcher-derived definitions (e.g., Langland, Lewis-Palmer, & Sugai, 1998). We define respect as a distinct construct, in that respect serves as a foundation to successful student-teacher and student-student relationships, which in turn forms a critical component to being connected to school and thus being ready and able to learn (Chafouleas & Miller, 2015). This also reflects the distinction made between compliance and social skills, on-task skills, and communication skills in the keystone behavior framework (Ducharme & Shecter, 2011). In this regard, while one can perceive possible overlap between disrespect and disruptive behavior, this study is consistent with the SWPBIS literature in conceptualizing defiance/disrespect as distinct from disruptive behavior. In fact, in a study of over 3000 schools collecting ODR data, defiance/disrespect was the most common reason for both minor and major ODR referrals at the elementary and middle school level, and far surpassed those issued for disruptive behavior (Gion, McIntosh, & Horner, 2014). Thus, we conceptualized these constructs as distinct in our approach, but were interested in empirically evaluating this assumption through implementation of the MTMM. Research has shown that positive relationships with adults and peers work as protective factors that strengthen resilience and ultimately student success, and respectful behavior functions to provide entry to and maintenance of positive relationships (Hamre & Pianta, 2001).

1.4.2. Assessment methods

The need to utilize multiple measures of behavior has been emphasized across the literature, with multimethod, multisetting, and multisource assessment acting a cornerstone of school psychological practice (Whitcomb & Merrell, 2013). With a focus on a single setting, the school, we review methods and sources as two potential sources of variance in ratings of the three key constructs of academically engaged, disruptive, and respectful behaviors.

1.4.2.1. Systematic direct observation. Systematic direct observation (SDO) encompasses a number of potential methods for collecting data on specific behaviors which share the common themes of utilizing systematic (as opposed to ad hoc) data collection through the direct observation of a target behavior (Suen & Ary, 1989). In practice, a researcher or educator may sit in the back of a classroom and, using a time-keeping device and a data recording sheet or application, indicate specific characteristics of a student's behavior. The specific behavior being observed is often guided by an operational definition, which specifies the topography of the behavior being observed in a clear, measureable fashion with examples and non-examples (Cooper, Heron, & Heward, 2007). This general observation method may be implemented using a wide range of tools, from standardized instruments such as the BOSS (Shapiro, 2011) to researcher- or practitioner-created measures which utilize methods like frequency counts, momentary time sampling, and partial interval recording. Surveys of school psychologists over the last two decades suggest that SDO-based methods remain one of the most popular tools for school-based behavior assessment (Shapiro & Heick, 2004; Wilson & Reschly, 1996).

1.4.2.2. Rating scales. In order to collect data relevant to summaries of student behavior, often related to a period of a few months, behavior rating scales may be used (Whitcomb & Merrell, 2013). These rating scales, such as the BASC-3 (Reynolds & Kamphaus, 2015) and the SSIS (Gresham & Elliott, 2008), are affective instruments which aggregate scores across multiple items relating to broad and narrow constructs to derive norm-referenced estimates of student behavior.

1.4.2.3. Direct behavior rating. Direct Behavior Rating, or DBR, is a method of behavior assessment wherein a student is observed for a period of time (e.g., during math class) and a retrospective rating of that student's behavior is made at the end of the observation period (Chafouleas, 2011). This rating corresponds to an estimate of the amount of time the student was engaged in a specific behavior during the observation period. Research on DBR has particularly focused on single-item scales (DBR-SIS) for use in progress monitoring (e.g., Chafouleas, Sanetti, Kilgus, & Maggin, 2012) and screening (e.g., Johnson et al., 2016), with a focus on collecting data related to academically engaged, disruptive, and respectful behaviors.

1.4.3. Raters

With most current assessment methods, tools for measuring behavior are completed by an individual, who considers a metric along which to judge a behavior and applies a subsequent rating to that behavior using that metric. In the context of SDO, this may take the form of a research assistant completing training in the operational definition of a behavior and then applying that definition to observed instances of behavior in real time. With rating scales, a parent or teacher may read an instrument's directions, view an item stem, and indicate the degree to which they perceive their child as engaging in the behavior indicated by that item. In any case involving human-mediated measurement, judgments are made and therefore may be expected to vary.

As comprehensively reviewed by De Los Reyes and Kazdin (2005), discrepancies among ratings of child psychopathology by different raters are a robust and consistent finding within the literature. Differences among ratings have been conceptualized within multiple frameworks; response bias may influence raters as they perceive a child as generally positive or negative (halo effect), or the rater may generally utilize the endpoints (leniency/severity) or midpoints (central tendency effect) of a scale (Whitcomb & Merrell, 2012). Differences between raters may result from actual differences in settings (Kazdin, 1979) and the contexts within which that behavior occurs and is interpreted by a rater (De Los Reyes, Henry, Tolan, & Wakschlag, 2009).

1.4.4. MTMM in behavior assessment

A review of the published literature for MTMM studies applying to child behavior in school-based settings revealed three previously-completed studies which reflect similar constructs to those examined in the current study (i.e., Efstratopoulou, Janssen, & Simons, 2012; Roberts, Milich, Loney, & Caputo, 1981; Spilt, Koomen, Stoel, Thijs, & van der Leij, 2011). Other applications of MTMM analyses to school-relevant behavior, but which target internalizing or more traditionally “affective” characteristics, have been conducted for constructs including depression and negative affect (e.g., Saylor, Finch, Furey, Baskin, & Kelly, 1984; Wolfe et al., 1987), self-concept (e.g., Marsh, Smith, & Barnes, 1983), student competence (Cole, Gondoli, & Peeke, 1998), and inter-relationships among these constructs (e.g., competence and depression; Cole, Martin, Powers, & Truglio, 1996), or examine relevant constructs outside the context of school-based settings (e.g., Hartung, McCarthy, Milich, & Martin, 2005).

The relationships between three measures of physical aggression and relational (or “nonaggressive antisocial”) aggression were examined by Spilt et al. (2011) with a Dutch sample of 117 kindergarten students. Results suggested a non-trivial method effect for interview- and observationally-derived data, although teacher-completed rating scales demonstrated desirable properties with significant path coefficients between traits and observed teacher-completed rating scales ($p < 0.001$) that were larger than those between the teacher source and observed rating scale data.

In an evaluation of aggressive as well as hyperactive and inattentive behavior, Roberts et al. (1981) utilized data from three teacher-completed rating scales pertaining to a total of 120 boys. Results from the three-method, three-trait matrix generally supported the convergent validity of each measure, although some evidence of a method effect was observed for the Conners Teachers' Rating Scale (Conners, 1997).

Utilizing reports from parents, classroom teachers, and physical education teachers, Efstratopoulou et al. (2012) examined the inter-relationships among data from a rating scale of motor behavior and two rating scales from the Achenbach System of Empirically Based Assessment. Data were collected pertaining to a randomly-selected sample of 841 elementary-aged students in Greece, with results suggesting that convergent validity was generally observed for most measure/trait pairs. Substantial correlation coefficients were observed within mono-method hetero-trait triangles, however, also suggesting the presence of significant method effects. Data derived from teacher- and parent-completed rating scales from the Achenbach System were also examined in Greenbaum, Dedrick, Prange, and Friedman (1994) and Stanger and Lewis (1993), with the addition of student-completed rating scales in the former study.

Although each of these MTMM-based investigations provides information regarding the validity and reliability of distinct sources of data pertaining to school-based student behavior, additional research pertaining to methods and traits commonly examined in school-based contexts is warranted. Most of the studies reviewed above relied exclusively on rating scales; observational measures were utilized by Spilt et al. (2011), but these were based on frequency counts of specifically-defined types of aggressive behavior, rather than the more-globally defined behaviors that may be more commonly assessed by school-based professionals. No identified studies have utilized MTMM analyses to assess the construct validity of globally-defined behaviors, measured using three distinct measures, with multiple raters in a school setting.

1.5. Current study

Despite the availability of the MTMM analytic method for examining the construct validity of data derived from common school-based behavior assessment tools, research is lacking pertaining to the inter-relationships among methods and traits derived from both (a) a wider range of behavior assessment methods and (b) focusing on a wider range of behaviors which may be target of school-based assessment teams. The current study aims to evaluate evidence of the convergent and discriminant validity of data derived from three distinct measures, using three distinct rater types, in order to measure three constructs which are theorized to hold substantial importance for students in schools. The following four research questions were explored through MTMM analyses with single-level and multi-level structural equation modeling.

1. To what extent are data from measures of school-based behavior associated?
   a. Based on the method (e.g., Spilt et al., 2011) and rater (e.g., De Los Reyes & Kazdin, 2005) effects generally observed in prior research, as well as method-specific studies suggesting substantial inter-relationships (e.g., DBR: Riley-Tillman, Chafouleas, Sassu, Chanese, & Glazer, 2008), it was hypothesized that significant correlations would be observed between measure-derived data.
2. How much of the variation in data from obtained measures is attributable to the traits (behaviors) relative to method?
   a. Although this question was considered exploratory in nature, based upon prior research (e.g., Efstratopoulou et al., 2012; Roberts et al., 1981) it was hypothesized that method effects would be observed.
3. Did the multilevel modeling approach utilized help to explain the variance in data from observed measures more effectively across trait and methods?
   a. It was hypothesized that, given the nested structure of the data, a class effect would be observed and when accounted for, would better explain the variance in data from observed measures across traits and methods (see Snijders & Bosker, 1999).
4. After controlling for effects of classes, did the relationships between trait factors and the relationships between method factors change?
   a. It was hypothesized that, after accounting for the effects of classes, the relationships between trait and method factors would change (see Snijders & Bosker, 1999).

2. Method

2.1. Participants and setting

Data were collected during the 2012–2013 school year in public school settings located within the northeastern and central regions of the United States. All data collection procedures were conducted in accordance with university-based human subjects review board approval. At each research site, convenience sampling was utilized wherein the primary investigator contacted key stakeholders (district-level and building-level administrators) to determine interest in participation. Subsequent to securing building-level administrator support for the project, teachers of grades 3–8 were recruited for participation. At each school, a brief informational presentation was delivered to teachers describing the nature and scope of the study, and teachers were asked to provide their contact information if they were interested in participating. As an incentive for participation, teachers were provided with a $50 gift card for time spent completing measures for this study. Within each teacher's class, parental consent forms were distributed, and students who returned signed consent forms and provided oral assent were deemed eligible for participation. Twelve school districts were represented in the study, in addition to one charter school and one large magnet school district. Participating teachers hailed from 19 different schools that varied considerably in regard to sociodemographic characteristics. In particular, free and reduced lunch rates ranged from 17 to 90% by school (mean = 45%), and the proportion of non-White students ranged from 3 to 99% (mean = 38%). In sum, 139 teachers were initially recruited for participation, corresponding to a total of 956 students. Of those initially recruited, 10 teachers did not complete the study (7%), leaving 129 teachers represented within the analytic sample. There were no significant differences with regard to demographic characteristics (race, gender, ethnicity, grade level taught) of the

Table 1
Student demographic characteristics.

Student characteristic	*n*	%
Gender		
Male	380	46
Female	451	54
Race		
White	542	65
African American	132	16
Multi-racial	82	10
Asian	33	4
Other	42	5
Ethnicity		
Hispanic	90	11
Non-Hispanic	741	89
Grade		
Upper elementary (3–5)	483	58
Secondary (6–8)	348	42
Special education status		
Identified with a disability	122	15
ELL status		
English language learner	30	4

recruited sample of teachers compared to the teachers who completed the study. Similarly, 956 students were initially recruited for participation, but 125 were excluded due to the teacher dropping out or absences that precluded use of the data (13%). Specifically, if a student did not have at least six DBR-SIS observation ratings and two SDO observation ratings, they were excluded from subsequent analyses in order to ensure adequate reliability of the data. Thus, 831 students comprised the analytic sample, all of whom had complete data. There were no significant differences with regard to demographic characteristics (race, gender, ethnicity, grade level) of the recruited sample of students compared to the students who comprised the analytic sample. In sum, the majority of participating students were White (65%), Non-Hispanic (89%), Upper-Elementary (3–5) grade students (58%). Each participating teacher had between 2 and 10 students participating in his/her class, but the majority of teachers (73%) had five or more students participating in his/her class. Participating teachers were predominately White (95%), Non-Hispanic (98%) and Female (84%). Student and teacher demographic characteristics are displayed in Tables 1 and 2, respectively.

2.2. Measures

2.2.1. DBR-SIS

DBR-SIS ratings reflect a rater's observation of the proportion of time a student is observed to be engaged in a target behavior on a 0 (never) to 10 (always) scale. Percentage of time anchors (0%, 50%, and 100%) are included on the beginning, middle, and end of the scale to aid in more accurate estimates of duration (Miller, Riley-Tillman, Chafouleas, & Schardt, 2017). DBR-SIS ratings were obtained on three target behaviors: academically engaged (AE), disruptive (DB), and respectful (RS). AE behavior was defined as

Table 2
Teacher demographic characteristics.

Teacher characteristic	*n*	%
Gender		
Male	21	16
Female	108	84
Race		
White	123	95
African American	2	2
Multi-racial	2	2
Other	2	2
Ethnicity		
Hispanic	2	2
Non-Hispanic	127	98
Grade		
Upper elementary (3–5)	66	51
Secondary (6–8)	63	49
Years teaching		
1–5	33	26
6–10	39	30
11 +	57	44

actively or passively participating in the classroom activity (e.g., writing, raising hand, looking at instructional materials). DB behavior was defined as student action that interrupts regular school or classroom activity (e.g., out of seat, playing with objects, talking/yelling about things that are unrelated to classroom instruction). RS behavior was defined as compliant and polite behavior in response to adult direction and/or interactions with peers and adults (e.g., follows teacher direction, pro-social interaction with peers, positive response to adult request). Studies on the psychometric properties of DBR-SIS with these three behaviors have provided support for the reliability of scores obtained and evidence of concurrent validity with systematic direct observations and behavior rating scales (Chafouleas, Sanetti, Jaffrey, & Fallon, 2012a, 2012b, 2012c; Johnson et al., 2016; Miller et al., 2015; Riley-Tillman et al., 2008). In particular, one-way intraclass correlation coefficients (ICCs) that examined variability between students and within observations all exceeded 0.90, suggesting that most of the variability in ratings were associated with the object of measurement as opposed to residual error (Johnson et al., 2016). Additionally, Miller et al. (2015) found strong correlations (0.52–0.68) between DBR-SIS and two widely used behavior rating scales: the BASC -2 Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007) and Social Skills Improvement System – Performance Screening Guide (SSiS-PSG; Elliott & Gresham, 2007). Paper forms were utilized for teacher- and student-completed DBR-SIS ratings.

2.2.1.1. Teacher report. DBR-SIS ratings were completed by teachers on participating students in their classroom twice per day. Research assistants worked with teachers to determine the target observation rating period as described in the procedures below. Immediately following the observation period, teachers completed DBR-SIS ratings for participating students. DBR-SIS teacher ratings were structured such that up to five students were rated by teachers on all three target behaviors twice daily across five days. For teachers with more than five participating students, upon completion of the first group of DBR-SIS ratings, the teacher subsequently rated the second group of students for five days. When more than five students were to be rated in a classroom, students were randomly assigned to the first or second rating group. Thus, a maximum of 10 teacher ratings were obtained for each participating student.

2.2.1.2. Student report. Participating students completed self-report DBR-SIS ratings following the same process as teachers (i.e., twice daily, immediately following the rating period). Students completed DBR-SIS ratings over a period of 2 consecutive weeks (10 school days). Thus, a maximum of 20 student report ratings were obtained for each participating student.

2.2.2. SDO

SDOs were conducted by trained research assistants during the same observation rating period as DBR collection. Training consisted of attending an informational session and completing practice observations using twelve 10 min video clips of classrooms that were master coded by two experts in behavior observation (one university-based researcher and one school-based practitioner), observing five students in sequential order with momentary time-sampling and 10 s intervals. These practice observation procedures were identical to those to be used by research assistants during study observations, with the exception of the observation length which was 10 min as opposed to 20. Both experts completed the online training module for DBR-SIS, which aligned to the specific behaviors observed in this study, prior to engaging in master coding (Chafouleas, Riley-Tillman, Jaffery, Miller, & Harrison, 2015). Trainees were required to meet or exceed a criterion of 90% agreement with the master code developed by the experts across three consecutive video clips for all three target behaviors in order to be eligible to conduct observations.

An observation protocol was developed using the same target behaviors and definitions included on the DBR-SIS form (AE, DB, RS). Momentary time sampling procedures were utilized, which are generally expected to produce more accurate estimates of duration than partial interval or whole interval recording procedures, depending on the characteristics of the specific behavior (Suen & Ary, 1989). Observers were cued at 10-second intervals using an audio track on an MP3 player. At the sound of the tone (i.e., the beginning of the interval), observers indicated whether the student demonstrated AE, DB, and RS behaviors. Up to five students were observed during a single observation session, with the observer rotating through each target student at each interval. Observers were instructed to conduct three to five 20-minute observations for each group of five students. At the end of the observation period, the overall percentage of the observation period that each student was engaged in each target behavior was calculated by dividing the number of intervals the behavior was present by the total number of intervals observed for each student. Inter-observer agreement (IOA) data were collected on 30% of observations.

2.2.3. BASC-2 behavior subscales (Reynolds & Kamphaus, 2004)

Behavior rating scales were selected to align with the traits of interest (AE, DB, RS). Although scales could readily be identified to correspond to AE and DB, the construct of RS has not been traditionally included in rating scales. Consequently, rating scales were only used in the measurement of AE and DB. To this end, the Attention Problems subscale and Hyperactivity subscale of the BASC-2 (Reynolds & Kamphaus, 2004) were utilized to evaluate AE and DB respectively. The Attention Problems (AP) subscale measures a student's inability to maintain attention and the tendency to be easily distracted from tasks requiring attention. While the AP subscale is scaled negatively as opposed to positively, the items were deemed to align with the construct of AE (e.g., is easily distracted from class work, listens carefully, pays attention, listens to directions). The Hyperactivity (H) subscale measures the tendency to be overly active, have poor self-control, and act without thinking. The items on the H subscale were deemed to align with the construct of DB (e.g., disrupts the schoolwork of other children, disrupts other children's activities, bothers other children when they are working, interrupts others). Combined sex norms were used to calculate *T*-scores for each subscale.

2.2.3.1. Teacher report. The teacher report form of the BASC-2 (Reynolds & Kamphaus, 2004) has two versions: one for children aged

6–11 and one for adolescents aged 12–21. The teacher completed the corresponding form based upon the student's age. For both the child and adolescent versions, the Attention Problems subscale consists of 7 items and the Hyperactivity subscale consists of 11 items. For each question, teachers respond using a 4-point Likert scale with the following response options: *Never*, *Sometimes*, *Often*, *Almost Always*. As reported in the technical manual, coefficient alpha for the norm sample ranged from 0.90–0.95 for the Attention Problems subscale, and from 0.91–0.95 for the Hyperactivity subscale. The subscales were also reported to have strong correlations (> 0.65) with similar subscales on the Conners' Teacher Rating Scale – Revised.

2.2.3.2. Student report. Like the teacher-report form, individual students completed the corresponding age-appropriate version of the self-report BASC-2 (Reynolds & Kamphaus, 2004) form (ages 8–11 or 12–21). For both the child and adolescent versions, the Attention Problems subscale consists of 9 items; the Hyperactivity subscale consists of 8 items for the child version and 7 items for the adolescent version. Both versions contain four True/False items, with the remaining items using a 4-point Likert scale with the following response options: *Never*, *Sometimes*, *Often*, *Almost Always*. As reported in the technical manual, coefficient alpha for the norm sample ranged from 0.76–0.79 for the Attention Problems subscale, and from 0.74–0.76 for the Hyperactivity subscale. The subscales were also reported to have strong correlations (> 0.55) with similar subscales on the Conners-Wells' Adolescent Self-Report Scale (Conners, 1997).

2.3. Procedures

During the fall of 2012, participant recruitment began across data collection sites. After securing administrator support, teacher permission to participate, and consent/assent from participating students and parents/guardians, data collection began in 2013. First, each classroom of participating teachers and students attended a class-wide hour-long training session on DBR-SIS. The classroom training session was delivered by trained graduate research assistants and was modeled after an online training module that was developed for DBR-SIS (see Chafouleas et al., 2015). A training protocol was developed for the graduate research assistants, which included a standardized PowerPoint presentation and a video modeling an exemplary training session, in order to ensure that standardized training was performed across sites.

The classroom training included (a) a description of DBR-SIS, (b) operational definitions of target behaviors (c) procedures for conducting ratings, (d) examples of how to assign ratings, and (e) opportunities for practice assigning ratings and feedback using brief video clips of a mock classroom. The DBR-SIS rating period varied depending on the students' grade level. That is, while elementary students have the same teacher all day, middle school students do not. Therefore, students in elementary grades (grades 3–5) provided ratings for the first half (e.g., before lunch) and second half (e.g., after lunch) of the day. For students in the middle school grades (grades 6–8), ratings were provided for the first half and second half of their participating teacher's class period. Systematic direct observations were conducted during DBR-SIS rating periods by trained external observers.

Teachers and students also completed BASC-2 (Reynolds & Kamphaus, 2004) subscale ratings. Assessment order (DBR-SIS or BASC-2 subscales) was counterbalanced across classrooms to control for potential order effects. Classrooms (i.e., students and teachers) were randomly assigned to counterbalancing conditions, with corrections made after random assignment to ensure even distribution of conditions within site and grade. One of two assessment order conditions were assigned: Completion of (a) DBR-SIS then BASC-2 subscales or (b) BASC-2 subscales then DBR-SIS. Demographic information pertaining to participating students and teachers was collected prior to the conclusion of the study.

2.4. Data analytic plan

2.4.1. Data preparation

Data were initially entered into a database in Microsoft Access, and a series of data verification and cleaning processes were implemented in SPSS Version 22 upon the completion of data collection (e.g., using syntax to ensure data were within permissible ranges, check for duplicate entries, and check for missing values). Moreover, student- and teacher-level data were double-entered for 30% of all teacher participants and compared against original data to check for any inconsistencies in scores at the individual item level. For example, a total of 7701 cases were entered for DBR-SIS IOA across student and teacher informants, and of those, 8% were flagged for potential inconsistencies between the original data file and IOA file. When inconsistencies were identified, the original paper forms were checked to determine the correct value and modify the data files. Student data were included in the final analysis file if they possessed non-missing values for all teacher- and self-reported rating scales, at least six observations for DBR-SIS ratings of AE, DB, and RS, and at least two SDO ratings, each of which needed to depict ratings for at least three minutes of total student-specific observation time (i.e., at least 18 intervals observed for that specific student in each included observation period). Although 956 students were initially recruited for participation, 65 could not participate due to the teacher dropping out, and 60 had absences that precluded use of the data (e.g., insufficient observations to provide a reliable estimate of student behavior). Thus, 831 students comprised the analytic sample, all of whom had complete data.

Prior to analysis, data were also transformed. First, all scores for DB and scores for the AP subscale were reverse-coded, to ensure that higher values indicated less risk across all three behaviors and thus facilitate ease of interpretation across behaviors. Second, all scores were transformed to align to a common metric across methods. Specifically, all scores were converted to a 0–10 scale prior to analysis. For SDO and *T*-scores, this simply required dividing the original value obtained by 10.

Table 3
MTMM matrix (3 Methods by 3 Traits).

MTMM matrix (3 × 3)												
Method				Student DBR-SIS			Teacher DBR-SIS			SDO		
	Trait	M	SD	AE	DB	RS	AE	DB	RS	AE	DB	RS
Student DBR-SIS	AE	8.88	1.15	**.92**[a]								
	DB	9.02	1.14	.62**	**.91**[a]							
	RS	9.09	1.07	.77**	.70**	**.92**[a]						
Teacher DBR-SIS	AE	8.52	1.36	.39**	.37**	.37**	**.93**[a]					
	DB	8.88	1.33	.33**	.39**	.39**	.77**	**.91**[a]				
	RS	9.22	1.12	.32**	.36**	.39**	.73**	.80**	**.92**[a]			
SDO	AE	8.70	1.20	.22**	.27**	.25**	.33**	.33**	.27**	**.93**[b]		
	DB	9.59	0.66	.16**	.19**	.19**	.24**	.27**	.23**	.77**	**.98**[b]	
	RS	9.92	0.30	.11**	.14**	.16**	.16**	.20**	.22**	.45**	.55**	**.99**[b]

Note. AE = academic engagement, DB = disruptive, RS = respectful, [a]denotes ICC reliability estimates, [b]denotes PABAK reliability estimates, **correlation is significant at 0.01 level. Hetrotrait-monomethod triangles are enclosed by a solid line, while heterotrait-heteromethod triangles are enclosed by a dashed line.

2.4.2. Data analyses

Each measure included within an MTMM is an observation on a target trait from a given method. To investigate effects of traits and methods on measures, the present study applied the standard confirmatory factor analysis (CFA) model of correlated traits and correlated methods to the MTMM, in which each measure simultaneously loads on its trait and method latent factors. In the standard CFA model of MTMM, trait factors are correlated with each other, as are the method factors. However, trait factors and method factors are usually assumed to be independent (Kenny & Kashy, 1992). In other words, the correlations between trait factors and method factors are constrained at zero. Theoretically, trait and method factors are supposed to explain variance components in measures that are exclusive from each other. Technically, allowing all trait and method factors to be correlated with each other will result in under-identified models. Empirically, identification issues are generally the major challenge for standard CFA models of MTMM data (Marsh & Bailey, 1991). Given the data used in the present study, the standard CFA model of the 3 × 3 MTMM (see

Table 4
MTMM matrix (5 Method by 2 Traits).

MTMM Matrix (5x2)													
				Student *T*-Score		Teacher *T*-Score		Student DBR-SIS		Teacher DBR-SIS		SDO	
	Trait	M	SD	AE	DB	AE	DB	AE	DB	AE	DB	AE	DB
Student *T*-score	AE	5.13	.93	**.76**[a]									
	DB	5.26	.95	.59**	**.78**[a]								
Teacher *T*-score	AE	5.02	1.02	.42**	.28**	**.87**[a]							
	DB	4.92	1.06	.33**	.37**	.72**	**.94**[a]						
Student DBR-SIS	AE	8.88	1.15	.42**	.28**	.33**	.28**	**.92**[b]					
	DB	9.02	1.14	.38**	.32**	.33**	.34**	.62**	**.91**[b]				
Teacher DBR-SIS	AE	8.52	1.36	.40**	.24**	.68**	.55**	.39**	.37**	**.93**[b]			
	DB	8.88	1.33	.35**	.28**	.56**	.63**	.33**	.39**	.77**	**.91**[b]		
SDO	AE	8.70	1.20	.18**	.16**	.23**	.27**	.22**	.27**	.33**	.33**	**.93**[c]	
	DB	9.59	.66	.15**	.13**	.20**	.25**	.16**	.19**	.24**	.27**	.77**	**.98**[c]

Table 3) successfully converged; however, the standard CFA model of the 5 × 2 MTMM (see Table 4) did not.

In light of the fact that measures of students' behaviors were naturally nested within their classrooms, a two-level standard CFA model of the 3 × 3 MTMM was tested to partial out the effects of teachers or peer influence across classrooms. Both one-level and two-level standard CFA models of the 3 × 3 MTMM were analyzed using Mplus 7.31 (2015). To address concerns regarding negative skewness and leptokurtic distributions of the data used in the present study, maximum likelihood estimation with robust standard errors (MLR) was used by specifying ESTIMATOR = MLR in Mplus. MLR is a non-normality robust technique that empirically produces quite robust parameter estimates, very good standard errors (SEs), and very good chi-square tests of model fit as well (Muthén & Kaplan, 1985). Multiple fit indices were used to evaluate the models; model fit was considered acceptable when CFI ≥0.95, SRMR ≤0.08, and RMSEA ≤0.06 (Hu & Bentler, 1999).

3. Results

3.1. Descriptive results

Two MTMMs are presented in Table 3 (3 × 3) and Table 4 (5 × 2); each matrix allows for a more nuanced analysis of traits and methods respectively. Within the MTMM, the diagonal coefficients are reliabilities and the off-diagonal coefficients are the Pearson product-moment correlations between observed variables from target methods on target traits. Within the MTMM, correlations can be classified by traits and methods, which are, theoretically from strongest to weakest: mono-trait hetero-method correlations, hetero-trait mono-method correlations, and hetero-trait hetero-method correlations.

3.1.1. Reliability

First, it should be noted that statistics used to derive reliability estimates differ from method to method. For DBR-SIS, reliability was measured using indices appropriate for single raters (see Chafouleas et al., 2013). For SDO, observer agreement indices were deemed most appropriate as these data are generally considered “the bedrock upon which sound behavioral measurement rests” (Watkins & Pacheco, 2000, p. 206). Conversely, for rating scale methods (*T*-scores), internal consistency was deemed most appropriate given this metric's prominence in traditional assessment literature (Hogan, Benjamin, & Brezinski, 2000). Specifically, one-way intra-class correlations [ICC (1,k)] were calculated for Teacher DBR-SIS and Student DBR-SIS ratings, which provided evidence of the consistency across observations over time (Shrout & Fleiss, 1979). Because 6–10 DBR-SIS ratings were conducted by teachers for each student included in analyses, median k values were 10 (range 6–11) for the teacher DBR-SIS ICC estimates, whereas median k values were 18 (range 6–21) for student DBR-SIS ICC estimates given ratings across approximately two weeks. Prevalence and bias adjusted kappa (PABAK) estimates were calculated for SDO, which provided a chance-adjusted measure of inter-rater agreement which also accounted for the high levels of prevalence observed (Byrt, Bishop, & Carlin, 1993); and Cronbach's alpha estimates were calculated to represent the overall consistency of teacher and students' *T*-scores. Across all methods, reliability indices were generally strong, with the exception of student *T*-scores. The ICC(1,k) estimates were high (> 0.90) across methods and traits. The PABAK coefficients were extremely high for SDO DB (k = 0.98) and SDO RS scores (k = 0.99) and slightly lower for SDO AE (k = 0.93). The alpha coefficients of Teacher *T*-scores and Student *T*-scores suggested that Student *T*-scores (α = 0.76–0.78) were less reliable than Teacher *T*-scores (α = 0.87–0.94) were for both traits AE and DB. At the same time, AE *T*-scores were less reliable than DB *T*-scores were for both Teacher *T*-scores and Student *T*-scores.

3.1.2. Bivariate correlations

All of the correlations presented in Table 3 and Table 4 were significantly different from zero at the 0.01 level. As previously described, the mono-trait hetero-method correlations are theoretically expected to be the strongest, and the hetero-trait-hetero-method correlations are theoretically expected to be the weakest, which was not always true for the MTMM data obtained in the present study. Relatively speaking, the strongest group of correlations in the MTMMs presented in Table 3 (the 3 × 2 MTMM) and Table 4 (the 5 × 2 MTMM) were the hetero-trait mono-method correlations, which ranged from 0.45 and 0.80 within the 3 × 3 MTMM and from 0.59 and 0.77 within the 5 × 2 MTMM. This finding suggests the existence of strong method factors, which was supported by the findings of the Structural Equation Models (SEM) presented later. Although the hetero-trait mono-method correlations appeared to be the strongest among the three groups of correlations, correlations were not too strong (< 0.90) to discriminate the theoretically different traits from each other.

Next, considering the correlations in each hetero-method block, there were different patterns observed between the 3 × 3 MTMM and 5 × 2 MTMM, and across traits and methods within either MTMM. Within the 3 × 3 MTMM, the mono-trait hetero-method correlations were relatively stronger than the hetero-trait hetero-method correlations in the Teacher DBR-SIS and Student DBR-SIS block; however, this finding did not hold in the Teacher DBR-SIS and SDO block nor the Student DBR-SIS and SDO block. Further, the mono-trait-hetero-method correlations between Teacher DBR-SIS scores and Student DBR-SIS scores were relatively stronger than the mono-trait-hetero-method correlations between those two DBR methods and the SDO scores across the three target traits. This finding provided stronger evidence of convergent validity between Teacher DBR-SIS and Student DBR-SIS methods than evidence of convergent validity between those two DBR methods and SDO.

Within the 5 × 2 MTMM, the correlations in the hetero-method blocks between SDO and the other four methods were the weakest correlations compared with the ones in the same column and row. The correlations in the hetero-method blocks between Teacher *T*-score and Teacher DBR-SIS were the strongest compared with the other hetero-method correlations. The mono-trait-hetero-method correlations within the hetero-method block between Teacher *T*-score and Teacher DBR-SIS were relatively strong (0.68 for AE, 0.63

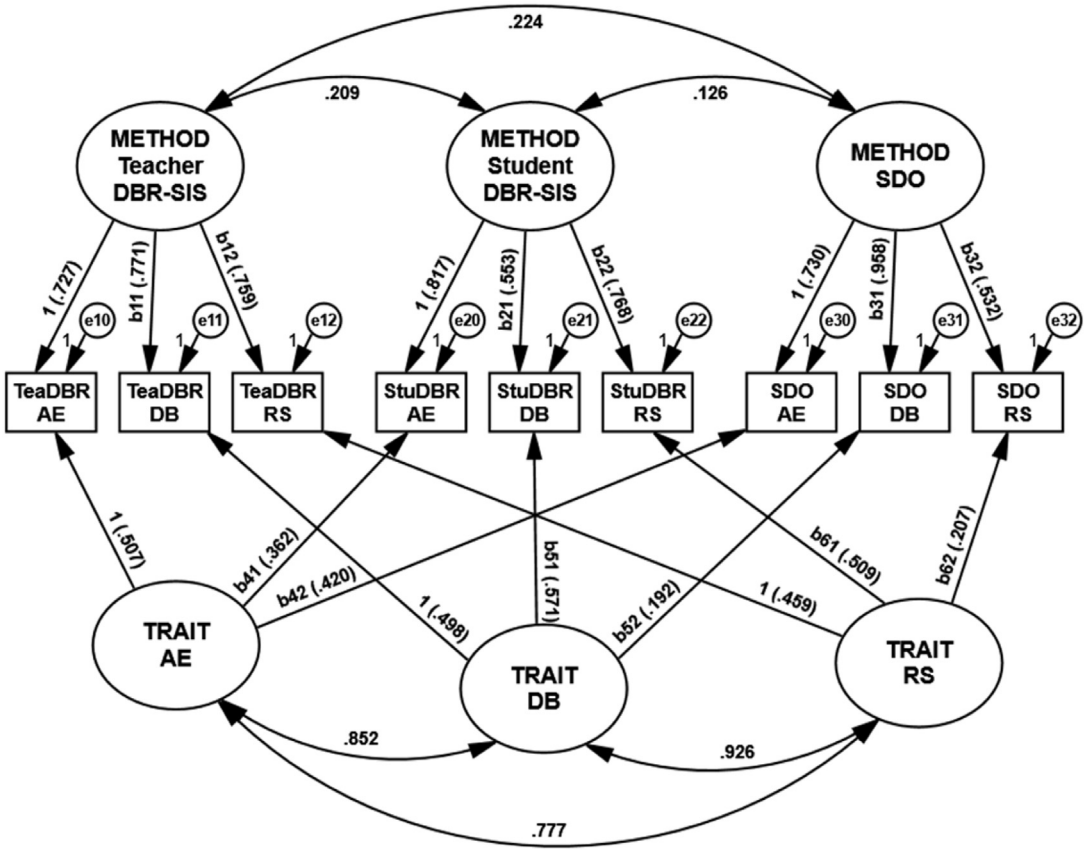

Fig. 1. Path diagram of the 1LMTMM model with correlations and standardized factor loadings (located in parentheses).
Note. AE stands for academically engaged, DB stands for disruptive behavior, RS stands for respectful behavior.

for DB) and substantially stronger than the corresponding correlations within the hetero-method block between Student *T*-score and Student DBR-SIS (0.42 for AE, 0.32 for DB). This finding indicated that teachers as raters provided more consistent evaluations on traits AE and DB than students as raters did. However, the hetero-trait-hetero-method correlations (0.56 and 0.55) within the hetero-method block between Teacher *T*-score and Teacher DBR-SIS were greater than all the correlations in the other hetero-method blocks. This finding suggested that the influence of teachers as raters was over and beyond the method, which did not hold for the influence of students as raters. This suggests that informant discrepancies may also vary as a function of the method used. This finding is of interest as it provides information regarding potential rater effects and method effects, which are often confounded in research examining informant discrepancies.

3.2. Model 1: One-level standard CFA model of the 3 × 3 MTMM

Model fit indices for the one-level standard CFA model of the 3 × 3 MTMM (hereafter labeled 1LMTMM model) indicated excellent model fit. The chi-square value (χ^2 = 18.959, df = 12, p = 0.09) was not statistically significant and the RMSEA = 0.027 (90% CI [0.000, 0.048]), CFI = 0.998, and SRMR = 0.013 all indicated that the 1LMTMM model fit the data very well. The path diagram of the 1LMTMM is presented in Fig. 1.

For any student i, there were nine observations on three traits (AE, DB, and RS) from three methods (Student DBR-SIS, Teacher DBR-SIS, and SDO). Then,

Table 5
1LMTMM model: Summary of the factor loadings, factor variance, residual variance, variance components explained by corresponding factors, r-square, and observed variance in each measure.

	Factor loadings						Variance component (VC)						R^2 $(1-\frac{Residual}{Total})$	Observed variance	Estimated intercept
	Factor	Method			Trait			Explained by method		Explained by trait		Residual	Total			
		T-DBR	S-DBR	SDO	AE	DB	RS									
	Variance	0.974	0.877	0.770	0.473	0.442	0.265	VC	%	VC	%					
T-DBR-AE		1.000	0.000	0.000	1.000	0.000	0.000	0.974	52.8%	0.473	25.7%	0.396	1.843	0.785	1.845	8.516
T-DBR-DB		1.043	0.000	0.000	0.000	1.000	0.000	1.060	59.5%	0.442	24.8%	0.280	1.782	0.843	1.772	8.884
T-DBR-RS		0.863	0.000	0.000	0.000	0.000	1.000	0.725	57.6%	0.265	21.0%	0.270	1.260	0.786	1.252	9.222
S-DBR-AE		0.000	1.000	0.000	0.604	0.000	0.000	0.877	66.7%	0.173	13.1%	0.265	1.315	0.798	1.320	8.876
S-DBR-DB		0.000	0.673	0.000	0.000	0.979	0.000	0.397	30.6%	0.424	32.6%	0.477	1.298	0.632	1.294	9.022
S-DBR-RS		0.000	0.879	0.000	0.000	0.000	1.059	0.678	59.0%	0.297	25.9%	0.174	1.149	0.849	1.143	9.088
SDO-AE		0.000	0.000	1.000	0.734	0.000	0.000	0.770	53.4%	0.255	17.7%	0.418	1.443	0.710	1.449	8.705
SDO-DB		0.000	0.000	0.717	0.000	0.189	0.000	0.396	91.9%	0.016	3.7%	0.019^{NS}	0.431	0.956	0.433	9.587
SDO-RS		0.000	0.000	0.181	0.000	0.000	0.120	0.025	28.3%	0.004	4.3%	0.060	0.089	0.326	0.089	9.916

Note. T-DBR stands for Teacher DBR-SIS, S-DBR stands for Student DBR-SIS, SDO stands for systematic direct observation, AE stands for academically engaged, DB stands for disruptive behavior, RS stands for respectful behavior, ^{NS}stands for not statistically significantly different from zero.

$$\begin{bmatrix} T-DBR-AE \\ T-DBR-DB \\ T-DBR-RS \\ S-DBR-AE \\ S-DBR-DB \\ S-DBR-RS \\ SDO-AE \\ SDO-DB \\ SDO-RS \end{bmatrix}_i = B_0 + \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & b_{11} & 0 & 0 \\ 0 & 0 & 1 & b_{12} & 0 & 0 \\ b_{41} & 0 & 0 & 0 & 1 & 0 \\ 0 & b_{51} & 0 & 0 & b_{21} & 0 \\ 0 & 0 & b_{61} & 0 & b_{22} & 0 \\ b_{42} & 0 & 0 & 0 & 0 & 1 \\ 0 & b_{52} & 0 & 0 & 0 & b_{31} \\ 0 & 0 & b_{62} & 0 & 0 & b_{32} \end{bmatrix} \begin{bmatrix} AE \\ DB \\ RS \\ T-DBR \\ S-DBR \\ SDO \end{bmatrix}_i + E_i$$

where B_0 and E_i are both 9×1 vectors. B_0 stands for intercepts of the nine measures, which are actually estimated averages of the nine measures after controlling for the effects of trait and method factors; E_i stands for random errors in the nine measures of student i, which vary across students. Taking measures "*T-DBR-AE*" and "*SDO-RS*" as examples,

$$T-DBR-AE_i = b_{01} + 1 \times AE + 1 \times T-DBR + e_{1i}$$

$$SDO-RS_i = b_{09} + b_{62} \times AE + b_{32} \times T-DBR + e_{9i}$$

So the variance in those two measures, *T-DBR-AE* and *SDO-RS*, across students is:

$$variance(T-DBR-AE_i) = 1^2 \times var(AE) + 1^2 \times var(T-DBR) + cov(AE, T-DBR) + var(e_{1i})$$

$$variance(SDO-RS_i) = {b_{62}}^2 \times var(RS) + {b_{32}}^2 \times var(SDO) + cov(RS, SDO) + var(e_{9i})$$

In the standard CFA model of MTMM data, the correlations between trait and method factors were constrained at zero, and therefore the $cov(AE, T-DBR)$ and $cov(RS, SDO)$ in the equations above were equal to zero. The measure $T-DBR-AE_i$ served as the marker variable for both its trait factor (AE) and its method factor (T-DBR), so loadings on those two factors were both constrained at 1. That is to say, $variance(T-DBR-AE_i)$ was the sum of its trait factor variance, its method factor variance, and its error variance. For $SDO-RS_i$, the loadings on its trait factor (RS) and its method factor (SDO) were both freely estimated, which are denoted as b_{62} and b_{32}. Therefore, $variance(SDO-RS_i)$ was the sum of its trait factor variance and its method factor variance weighted by its loadings on those factors plus its error variance. In each measure, the size of the variance component explained by the trait factor or the method factor was determined by the absolute value of loadings and the size of factor variance.

In Table 5, summarized factor loadings, factor variance, residual variance, r-square, and the calculation results of variance components explained by corresponding factors are presented. In general, the 1LMTMM model explained a large proportion of the variance in the nine measures except for that corresponding to trait RS from the SDO method (labeled SDO-RS). Only 32.6% of the variance in the SDO-RS was explained by the 1LMTMM model, among which only 4.3% of the variance in the SDO-RS was explained by the trait factor (RS). For the other eight measures, the percentages of variance explained by the 1LMTMM model ranged between 63.2% and 95.6%. Comparing the percentages of variance in each measure explained by method factors and trait factors with each other, it is apparent that method factors had stronger effects on measures than trait factors with one exception, which was the measure on trait DB from method Student DBR-SIS (labeled S-DBR-DB). For Student DBR-DB, the variance explained by the method factor (Student DBR-SIS factor, 0.397 corresponding to 30.6% of total variance) was slighter lower than the variance explained by the trait factor (DB, 0.424 corresponding to 32.6% of total variance). The strongest method effects were observed on trait DB from the method SDO (labeled SDO-DB), in which 91.9% of variance was explained by the method factor (SDO) but only 3.7% of the variance was explained by the trait factor (DB).

Next, the magnitude of correlations was examined using the interpretive guidelines discussed by Cohen (1988): correlations < 0.30 were considered weak, 0.30 to 0.49 were considered moderate, and 0.50 and above were considered strong. Method factors were weakly but significantly correlated with each other ($r_{T-DBR-SIS,\ S-DBR-SIS} = 0.29$, $r_{T-DBR-SIS,\ SDO} = 0.22$, $r_{S-DBR-SIS,\ SDO} = 0.13$); at the same time, the latent trait factors were strongly and significantly correlated with each other ($r_{AE,\ DB} = 0.85$, $r_{AE,\ RS} = 0.78$, $r_{DB,\ RS} = 0.93$). After accounting for the effects of trait factors, the correlations between method factors were weak, indicating a low degree of association between methods. The correlation between DB and RS was very strong, suggesting a high degree of association between these traits.

3.3. Model 2: Two-level standard CFA model of the 3 × 3 MTMM

The intra-class correlation (ICC) indicates how much variance in measures exists between classes. Given the MTMM data in the present study, respectively for traits AE, DB, and RS, the ICCs of the measures from the teacher DBR-SIS were 0.232, 0.244, and 0.318; the ICCs of the measures from the student DBR-SIS were 0.124, 0.053, and 0.069; and the ICCs of the measures from SDO were 0.26, 0.244, and 0.046. The magnitude of the majority of ICCs obtained (i.e., > 0.10; Kline, 2011) suggested the importance of running multilevel models to examine the effect of classes. Therefore, a two-level standard CFA model of the 3 × 3 MTMM (hereafter labeled 2LMTMM model) was tested. In brief, the 2LMTMM model fit the MTMM data quite well. For the sake of convergence, the between-classroom level residual variance in observed variables was constrained to be equal across traits from the same method. In addition, the correlations between the three method factors at the between-class level were not statistically significantly different from zero. By constraining correlations between the three method factors equal to zero at the between-class level, the Bayesian information

Table 6
2LMTMM model: Summary of the factor loadings, factor variance, residual variance, variance components explained by corresponding factors, r-square, and observed variance in each measure.

		Factor loadings						Variance component (VC)						R^2 $(1-\frac{\sigma^2}{\Sigma^2})$	
		Method			Trait			Explained by method		Explained by trait		Residual	Sub total		
		Factor	T-DBR	S-DBR	SDO	AE	DB	RS								
Within class		Variance	1.119	0.874	0.595	0.136^{NS}	0.434	0.298	VC	%	VC	%	σ^2	Σ^2		Total R^2
	T-DBR-AE		1.000	0.000	0.000	1.000	0.000	0.000	1.119	77.7%	0.136	9.4%	0.186^{NS}	1.441	0.871	
																0.870
	T-DBR-DB		0.807	0.000	0.000	0.000	1.000	0.000	0.729	53.0%	0.434	31.5%	0.213	1.376	0.845	
																0.843
	T-DBR-RS		0.628	0.000	0.000	0.000	0.000	1.000	0.441	49.9%	0.298	33.7%	0.145	0.884	0.836	
																0.850
	S-DBR-AE		0.000	1.000	0.000	0.129 ^{NS}	0.000	0.000	0.874	75.1%	0.002	0.2%	0.288	1.164	0.753	
																0.772
	S-DBR-DB		0.000	0.848	0.000	0.000	0.436	0.000	0.628	51.1%	0.083	6.7%	0.518	1.229	0.579	
																0.596
	S-DBR-RS		0.000	0.974	0.000	0.000	0.000	0.426	0.829	77.1%	0.054	5.0%	0.192	1.075	0.821	
																0.828
	SDO-AE		0.000	0.000	1.000	1.612 ^{NS}	0.000	0.000	0.595	54.2%	0.353	32.2%	0.149^{NS}	1.097	0.864	
																0.899
	SDO-DB		0.000	0.000	0.669	0.000	0.263	0.000	0.266	77.8%	0.030	8.8%	0.046 ^{NS}	0.342	0.866	
																0.901
	SDO-RS		0.000	0.000	0.235	0.000	0.000	0.120	0.033	38.1%	0.004	5.0%	0.049	0.086	0.431	
																0.417
Between class		Variance	0.294	0.102	0.247	0.055^{NS}	0.061^{NS}	0.081^{NS}	VC	%	VC	%				Intercepts
	T-DBR-AE		1.000	0.000	0.000	1.000	0.000	0.000	0.294	70.2%	0.055	13.1%	0.070	0.419	0.833	
																8.499
	T-DBR-DB		0.958	0.000	0.000	0.000	1.000	0.000	0.270	67.3%	0.061	15.2%	0.070	0.401	0.825	
																8.860
	T-DBR-RS		0.912	0.000	0.000	0.000	0.000	1.000	0.245	61.8%	0.081	20.5%	0.070	0.396	0.823	
																9.193
	S-DBR-AE		0.000	1.000	0.000	0.867	0.000	0.000	0.102	70.7%	0.041	28.6%	0.001	0.144	0.993	
																8.876
	S-DBR-DB		0.000	0.252^{NS}	0.000	0.000	0.937	0.000	0.006	10.6%	0.054	87.7%	0.001	0.061	0.984	
																9.027
	S-DBR-RS		0.000	0.415	0.000	0.000	0.000	0.751	0.018	27.3%	0.046	71.1%	0.001	0.064	0.984	
																9.092
	SDO-AE		0.000	0.000	1.000	1.397^{NS}	0.000	0.000	0.247	69.1%	0.107	30.0%	0.003	0.357	0.992	
																8.739
	SDO-DB		0.000	0.000	0.616	0.000	−0.065^{NS}	0.000	0.094	96.6%	0.000	0.3%	0.003	0.097	0.969	
																9.600
	SDO-RS		0.000	0.000	0.024^{NS}	0.000	0.000	−0.005^{NS}	0.000	4.5%	0.000	0.1%	0.003	0.003	0.046^{NS}	
																9.917

Note: T-DBR stands for Teacher DBR-SIS, S-DBR stands for Student DBR-SIS, SDO stands for systematic direct observation, AE stands for academically engaged, DB stands for disruptive behavior, RS stands for respectful behavior, ^{NS}stands for not statistically significantly different from zero. ^{c}indicates the parameter has been constrained at the given value. Total $R^2 = (\mathbf{1}-\frac{\boldsymbol{\sigma^2_w}+\boldsymbol{\sigma^2_b}}{\boldsymbol{\Sigma^2_w}+\boldsymbol{\Sigma^2_b}})$.

criterion (BIC) reduced from 15,792.394 to 15,781.955, which favored the more parsimonious model. Further, the CFI was high (0.994), SRMR-Within was low (0.013), whereas the SRMR-Between was higher (0.155), and the RMSEA was low (0.023). Overall, these data suggest that the 2LMTMM model fit the data very well.

The purpose of running the 2LMTMM model was to first account for the effects of classes on measures by accounting for variation in behavior across classes, and then to explore whether or not variance in measures could be better described by introducing the second level modeling. Compared with the percentages of variance explained by the latent factors in 1LMTMM (R^2 in Table 5), the percentages of variance (Total R^2 in Table 6) in measures explained by the within and between class latent factors in 2LMTMM increased for the measures on traits AE and RS from methods Teacher DBR-SIS and SDO.

The within-level correlations between latent method factors ($r_{WT-DBR-SIS,\ WS-DBR-SIS} = 0.48$, $r_{WT-DBR-SIS,\ WSDO} = 0.20$, $r_{WS-DBR-SIS,\ WSDO} = 0.28$) were still weak between SDO and DBR-SIS, but the correlation between Teacher DBR-SIS and Student DBR-SIS was moderate and notably stronger than the corresponding estimates obtained by the 1LMTMM model. The within-level correlations between trait factors were moderately strong or strong ($r_{WAE,\ WDB} = 0.70$, $r_{WAE,\ WRS} = 0.53$, $r_{WDB,\ WRS} = 0.90$), and the between-class level correlations were strong ($r_{BAE,\ BDB} = 0.88$, $r_{BAE,\ BRS} = 0.79$, $r_{BDB,\ BRS} = 0.86$). It was noticeable that the correlations between AE and RS were weaker than the other two correlations at either level. The stronger correlations between DB and the other two traits indicate a high degree of association between them.

4. Discussion

The purpose of this study was to examine the relation between traits and methods within the domain of behavior assessment, in an effort to better understand the nature of the school-based assessments as well as their relative strengths and weaknesses. With regard to psychometric defensibility, several interesting findings emerged. First, estimates of reliability, indices for each assessment method were generally high, with the exception of internal consistency estimates of Student *T*-scores. These findings are consistent with prior research (e.g., Achenbach, McConaughy, & Howell, 1987) which suggest that reliability estimates for child self-report measures tend to be lower than estimates obtained from adult raters of children, especially for ratings of externalizing behaviors. Although it is important to note that different reliability indices were calculated depending on the method, each estimate represents the consistency of scores obtained on the dimension of reliability deemed most salient to the specific method. Next, we turn to evidence for validity of scores, which were tied to specific research questions outlined.

4.1. Degree of association between behavioral assessment methods

With regard to our first research question exploring the extent to which measures of school-based behavior were associated, correlations presented within the MTMM yielded several interesting findings. Consistent with our hypothesis, significant correlations were observed between measures. The magnitude of the validity coefficients (mono-trait hetero-method correlations) varied, with most falling within the small to moderate range using the interpretive guidelines provided by Cohen (1988). Not surprisingly, validity coefficients were strongest between Teacher DBR-SIS and Teacher *T*-scores, as these coefficients shared the same adult informant. Conversely, the SDO method block consistently displayed the lowest validity coefficients with the other methods. It appears that SDO had the strongest relationship to Teacher DBR-SIS relative to the other methods, albeit a modest relationship. The strongest relationships in the matrices were generally observed for the hetero-trait-mono-method triangles, indicating the presence of strong methods factors. This finding was also confirmed through the SEM analyses performed. However, these findings must also be interpreted within the context of the various informants. In particular, these findings also suggest that teachers, students, and outside observers may have different perceptions of student behavior. Further, the magnitude of the hetero-trait hetero-method correlations were similar in magnitude to the validity coefficients, calling into question the extent to which the constructs of AE, DB, and RS are independent of each other. This finding was further highlighted within the SEM models and the strong correlations observed between the latent trait factors. In particular, correlations between DB and RS were very strong, and as previously reviewed, there is inconsistency regarding the extent to which researchers conceptualize these traits as distinct. This finding can be interpreted in several different ways. That is, to what extent are the behaviors actually distinct and are they better reflected as a broader construct, or are they in fact distinct but with a high degree of co-occurrence existing between these traits? It may be the case that many of these behaviors simply co-occur. As the first study of this nature, future research is needed to further elucidate the nature of such relations.

4.2. Trait versus method variance

Findings from the SEM models permitted a more detailed analysis of the variation in scores that may be attributed to traits relative to methods, which was germane to answer the second research question. As hypothesized, examination of the one-level model suggested that a greater proportion of the variance in scores was explained by method factors relative to trait factors. Indeed, examination of each measure individually revealed that the proportion of variance explained by the model that was attributable to method ranged from 28% to 92% whereas the proportion of variance attributable to trait ranged from 4% to 33%. Interestingly, the relation between methods was weaker than the relation between traits. That is, although the methods were weakly correlated with each other, the traits were strongly correlated with each other, with DB and RS exhibiting the strongest relation within classes. Interestingly, from a discriminant validity standpoint, the relation between AE and RS was weaker compared to AE and DB. Drawing from our earlier discussion of the keystone behavior framework (Ducharme & Shecter, 2011), many questions remain regarding the nature of such relations between traits. Within this framework, keystone behaviors are conceptualized as “relatively circumscribed”

but related to other responses. As an example, the authors state that noncompliance is prevalent in nearly all externalizing disorders in children. Similarly, it stands to reason that students who display low levels of respectful behavior would also display high levels of disruptive behavior. Examination of the magnitude of trait correlations in other MTMM studies yields some consistent findings. For example, Spilt et al. (2011) found the traits of physical aggression and nonaggressive antisocial behavior to be strongly correlated at 0.81; these traits are clearly distinct (one involves physical harm to others, while the other does not) but highly co-occur. Similarly, Hartung et al. (2005) found the traits of hyperactivity and conduct disorder to be strongly correlated at 0.57. Disentangling the nature of these relations is further complicated by the proportion of variance that was attributable to methods rather than traits. What are we actually capturing within each assessment? It is only through further study that we can begin to understand these issues.

With regard to methods, modest correlations between teacher rating methods and direct observations were also obtained in an MTMM study conducted by Spilt et al. (2011), where the magnitude of the relation was between 0.19 and 0.14. These findings suggest that the methods used to measure traits have significant impacts on the results obtained. However, given the nested structure of the data, multi-level modeling was necessary to account for the effects of classes. Accounting for the class effect is important due to potential differences in student behavior between classes, as well as teacher stringency or leniency in assigning ratings. Importantly, to our knowledge, no school-based MTMM study published to date has utilized multi-level modeling in the analysis of such data.

4.3. Multilevel model

Our final research questions addressed whether the multi-level modeling approach utilized was more effective in explaining variance components than the one-level model, and whether this approach impacted findings relative to the one-level model. Consistent with our hypotheses, the multilevel model was more effective in explaining variance components and differed from the one-level model. A greater proportion of variance was explained by the two-level model for two of the measures, and both relative and absolute fit indices indicated good model fit. The within-class portion of the two-level model was consistent with the one-level model, with a greater proportion of variance attributable to methods relative to traits. The between-class portion of the two-level model also indicated a greater proportion of variance attributable to method, with the exception of Student DBR-SIS ratings of DB and RS, where a substantial proportion of the variance was attributable to the traits.

4.4. Implications for practice

The analytic methods employed in this study permitted the evaluation of method and/or rater effects relative to the traits measured, and thus yield several interesting findings. Principally, nearly all of the method factors had stronger effects on measures than trait factors. Thus, it is essential to consider the influence of method when using these data for decision-making. This is particularly important given that all prevention and intervention efforts rely on psychometrically sound assessments to support early identification, intervention, and problem-solving. Thus, understanding the extent to which each method actually captures trait-level variance is critical.

Considering the methods used to estimate these traits, relative strengths and weaknesses are apparent. With regard to Teacher DBR-SIS, the variance attributable to traits was higher for DB and RS relative to AE, yet the opposite pattern was observed for SDO. That is, the SDO procedures utilized in this study appeared to capture trait-variance for AE much better than for DB or RS. In fact, the variance attributable to DB and RS was negligible using the SDO procedures utilized in this study. Therefore, Teacher DBR-SIS ratings may be particularly relevant to capturing variance associated with DB and RS whereas SDO procedures appear well-suited for AE relative to the other constructs. Notably, the SDO procedures for RS were particularly troublesome, with very little variance accounted for by either trait or method. This might be due to the low base rate of these behaviors and thus may not be captured in relatively short observation periods. With regard to Student DBR-SIS ratings, limited research has examined the psychometric defensibility of these ratings, which could potentially be efficient to collect and ease the burden of data collection on teachers. To this end, within-class estimates suggest that very little variance was attributable to the traits, yet between-class estimates suggest that a modest (AE) to large (DB and RS) proportion of the variance is attributable to these traits. Thus, although student DBR-SIS ratings may not capture trait variance particularly well within class, it may be better suited in decision-making between-classes. Although the influence of rater was not tested empirically through the SEM modeling, the 5×2 MTMM provides some interesting findings in this regard. In particular, evidence for possible rater effects were observed for teacher ratings of students' behavior, but not for student self-report ratings.

The findings from the present study further reinforce the importance of multi-method, multi-source assessment practice, given potential biases associated with any single method of measurement. Relying on only one method of assessment or one source of information may bias the results obtained, thus collecting corroborative data using multiple methods is advised (Shapiro & Kratochwill, 2000). Although low-inference data such as SDO have been considered the “gold standard” in regard to the objective measurement of behavior (Baer, Wolf, & Risley, 1987), these findings point to weaknesses in capturing trait-level variance even with this method. In particular, the variability in measures of SDO-RS and SDO-DB were quite small using the momentary time sampling procedures in this study. Epps (1985) argued intuitively that an observation system is valid only when it accurately measures what it purports to measure, thus these findings shed light on the extent to which trait-level variance is actually captured using differing behavior assessment methodologies. Therefore, it is essential that researchers and practitioners alike understand the strengths and limitations of their behavior assessment methods and utilize multiple data sources in order to triangulate findings and engage in more defensible data-based decision-making processes.

4.5. Limitations and directions for future research

Several limitations should be considered in the context of this study. First, measures of student behavior were obtained for students in grades 3–8, thus the extent to which these findings generalize to other age groups remains unknown and should be investigated in future research. Similarly, data were obtained from 2 to 10 students in each classroom, thus the extent to which findings generalize to whole classrooms requires further investigation. Second, specific time-sampling procedures were used to derive SDO estimates, thus the degree to which the findings generated using momentary time sampling procedures with 10-second intervals would compare to other time sampling procedures (i.e., partial interval or whole interval) and/or with alternative interval lengths is not known. Similarly, observations were restricted to only 20-minute intervals, thus encompassing a smaller observation period than DBR-SIS ratings. Although multiple observations were conducted in an effort to obtain a representative and reliable estimate of behavior, the extent to which the observations were representative remains unclear. Thus, the findings from this study represent only one SDO method, and future research is needed to determine how findings may differ utilizing other SDO methods (e.g., Johnson, Chafouleas, & Briesch, 2017). Similarly, the findings from this study represent only one rating scale method (BASC-2 subscales). Although an analysis of the items included in the AP and H subscales determined that they were well-aligned to the operational definitions of AE and DB respectively, they were not perfectly aligned. Additional research is also needed with more robust and diverse rating scale methods. Third, in an effort to evaluate constructs that may act as GOMs for student behavior, the traits of academic engagement, disruptive, and respectful behavior were included in the present investigation. However, additional constructs could be included in future investigations. That is, consensus has not been reached regarding: (a) which behavioral constructs best perform as GOMs, and (b) how the salience of these constructs may change with youth development. Thus, ample opportunities exist for future research to advance knowledge in this area.

An additional challenge within the context of the present study involved identifying a rating scale method to align with the construct of respectful behavior that could be used within the context of the MTMM. Despite the prominence of the construct of defiance/disrespect within the SWPBIS literature (e.g., Gion et al., 2014), an equivalent rating scale measure could not be identified for the purpose of this study. This challenge reflects the complexity with how behavioral constructs are conceptualized and measured. Indeed, Silva's (1993) analysis of the traditions of behaviorism and psychological testing developing in isolation have certainly influenced the state of behavioral assessment to date.

Finally, convergence problems are common when applying the standard CFA model to the MTMM. In the case of the 5×2 MTMM, Heywood cases contributed to convergence problems. We attempted to resolve these cases by adding additional constraints to the model, but in doing so the parameter estimates became untenable. Consequently, we restricted our analysis of the 5×2 only to descriptive analysis of the correlation matrix, since the standard CFA model could not be applied to the 5×2. Future research is needed to extend this work in order to better understand the relative contributions of traits and methods in school based behavioral assessments.

4.6. Conclusion

Despite longstanding calls for increased attention to reliability and validity evidence for behavioral assessment methods (e.g., Cone, 1977; Silva, 1993), surprisingly few MTMM studies have been conducted in the area of behavioral assessment in educational settings. Indeed, our search of such studies yielded only three published manuscripts (Efstratopoulou et al., 2012; Roberts et al., 1981; Spilt et al., 2011), none of which utilized multi-level modeling to account for the nested structure of the data. Consequently, this investigation represents a substantial contribution to the research literature regarding the amount of trait-level variance actually captured by various behavior assessment methods. However additional investigations are greatly needed to replicate and expand upon these findings.

The present study on school-based behavioral assessment methods was conducted in order to shed light on the extent to which behavior assessment methods capture intended trait-level variance and determine the degree to which methods utilized impact findings. Support was found for the presence of strong methods effects for all of the assessment methods utilized, which included three of the most common behavior assessment methods used in schools. These findings are consistent with long-standing recommendations calling for the use of multi-source multi-method assessments in the realm of behavioral assessment.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *The standards for educational and psychological testing.* Washington, DC: AERA.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20*, 313–327.

Bambara, L. M. (2005). Evolution of positive behavior support. In L. Bambara, & L. Kern (Eds.). *Individualized supports for students with problem behaviors: Designing positive behavior plans* (pp. 1–24). New York, NY: Guilford Press.

Bradley, R., Doolittle, J., & Bartolotta, R. (2008). Building on the data and adding to the discussion: The experiences and outcomes of students with emotional disturbance. *Journal of Behavioral Education, 17*, 4–23.

Bridgeland, J., Bruce, M., & Hariharan, A. (2013). The missing piece: A national teacher survey on how social and emotional learning can empower children and transform schools (Retrieved from) https://www.casel.org/wp-content/uploads/2016/01/the-missing-piece.pdf.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423–429.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Chafouleas, S. M. (2011). Direct behavior rating: A review of the issues and research in its development. *Education and Treatment of Children, 34*, 575–591.
Chafouleas, S. M., Kilgus, S. P., Jaffery, R., Riley-Tillman, T. C., Welsh, M., & Christ, T. J. (2013). Direct behavior rating as a school-based behavior screener for elementary and middle grades. *Journal of School Psychology, 51*, 367–385.
Chafouleas, S. M., & Miller, F. G. (2015). Respect. In W. George Scarlett (Ed.). *Encyclopedia of classroom management* (pp. 670–672). Thousand Oaks, CA: SAGE Publications.
Chafouleas, S. M., Riley-Tillman, T. C., Jaffery, R., Miller, F. G., & Harrison, S. E. (2015). Preliminary investigation of the impact of a web-based module on direct behavior rating accuracy. *School Mental Health, 7*, 92–104.
Chafouleas, S. M., Sanetti, L. M. H., Jaffrey, R., & Fallon, L. M. (2012a). An evaluation of a classwide intervention package involving self-management and a group contingency on classroom behavior of middle school students. *Journal of Behavioral Education, 21*, 34–57.
Chafouleas, S. M., Sanetti, L. M. H., Jaffrey, R., & Fallon, L. M. (2012b). An evaluation of a classwide intervention package involving self-management and a group contingency on classroom behavior of middle school students. *Journal of Behavioral Education, 21*, 34–57.
Chafouleas, S. M., Sanetti, L. M. H., Jaffrey, R., & Fallon, L. M. (2012c). An evaluation of a classwide intervention package involving self-management and a group contingency on classroom behavior of middle school students. *Journal of Behavioral Education, 21*, 34–57.
Chafouleas, S. M., Sanetti, L. M. H., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change across consultation cases using Direct Behavior Rating Single-Item Scales (DBR-SIS). *Exceptional Children, 78*, 491–505.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
Cole, D. A., Gondoli, D. M., & Peeke, L. G. (1998). Structure and validity of parent and teacher perceptions of children's competence: A multi-trait–multimethod–multigroup investigation. *Psychological Assessment, 10*, 241–249.
Cole, D. A., Martin, J. M., Powers, B., & Truglio, R. (1996). Modeling causal relations between academic and social competence and depression: A multitrait-multimethod longitudinal study of children. *Journal of Abnormal Psychology, 105*, 258–270.
Cone, J. D. (1976, September). Multitrait-multimethod matrices in behavioral assessment. *Paper presented at meeting of the American Psychological Association, Washington, DC*.
Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8*, 427–430.
Conners, C. K. (1997). *The Conners rating scales – REVISED.* North Towanda: Multi-health Systems.
Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.
Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.
De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology, 37*, 637–652.
De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483–509.
Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15*, 358–374.
DiPerna, J. C., & Elliott, S. N. (2002). Promoting academic enablers to improve student achievement: An introduction to the mini-series. *School Psychology Review, 31*, 293–297.
Ducharme, J. M., & Shecter, C. (2011). Bridging the gap between clinical and classroom intervention: Keystone approaches for students with challenging behavior. *School Psychology Review, 40*, 257–274.
Efstratopoulou, M., Janssen, R., & Simons, J. (2012). Agreement among physical educators, teachers and parents on children's behaviors: A multitrait-multimethod design approach. *Research in Developmental Disabilities, 33*, 1343–1351.
Elliott, S. N., & Gresham, F. M. (2007). *SSIS performance screening guide.* Minneapolis, MN: Pearson.
Epps, S. (1985). Best practices in behavioral observation. In A. Thomas, & J. Grimes (Eds.). *Best practices in school psychology* (pp. 95–111). Kent, OH: National Association of School Psychologists.
Fernet, C., Guay, F., Senécal, C., & Austin, S. (2012). Predicting intraindividual changes in teacher burnout: The role of perceived school environment and motivational factors. *Teaching and Teacher Education, 28*, 514–525.
Gion, C. M., McIntosh, K., & Horner, R. H. (2014). Patterns of minor office discipline referrals in schools using SWIS (Retrieved from) http://www.pbis.org/blueprint/evaluation-briefs/patterns-of-minor-odrs.
Greenbaum, P. E., & Dedrick, R. F. (1998). Hierarchical confirmatory factor analysis of the Child Behavior Checklist/4–18. *Psychological Assessment, 10*, 149–155.
Greenbaum, P. E., Dedrick, R. F., Prange, M. E., & Friedman, R. M. (1994). Parent, teacher, and child ratings of problem behaviors of youngsters with serious emotional disturbances. *Psychological Assessment, 6*, 141–148.
Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives in research and practice. *School Psychology Review, 31*, 328–349.
Gresham, F. M. (2015). *Disruptive behavior disorders: Evidence-based practice for assessment and intervention.* New York, NY: Guilford Press.
Gresham, F. M., Cook, C. R., Collins, T., Dart, E., Rasetshwane, K., Truelson, E., & Grant, S. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating System-Teacher Form. *School Psychology Review, 39*, 364–379.
Gresham, F. M., & Elliott, S. N. (2008). *Social skills improvement system: Rating scales.* Bloomington, MN: Pearson.
Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72*, 625–638.
Hartung, C. M., McCarthy, D. M., Milich, R., & Martin, C. A. (2005). Parent-adolescent agreement on disruptive behavior symptoms: A multitrait-multimethod model. *Journal of Psychopathology and Behavioral Assessment, 27*, 159–168.
Hofling, V., Schermelleh-Engel, K., & Moosbrugger, H. (2009). Analyzing multitrait-multimethod data: A comparison of three approaches. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5*, 99–111.
Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods. *Educational and Psychological Measurement, 60*, 523–531.
Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement.* New York, NY: Guilford.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
Institute of Medicine (2012). *An integrated framework for assessing the value of community-based prevention.* Washington, DC: The National Academies Press.
Johnson, A. H., Chafouleas, S. M., & Briesch, A. M. (2017). Dependability of data derived from time sampling methods with multiple observation targets. *School Psychology Quarterly, 32*, 22–34.
Johnson, A. H., Miller, F. G., Chafouleas, S. M., Welsh, M. E., Riley-Tillman, T. C., & Fabiano, G. A. (2016). Evaluating the technical adequacy of DBR-SIS in tri-annual behavioral screening: A multisite investigation. *Journal of School Psychology, 54*, 39–57.
Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.
Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2 behavioral and emotional screening system.* Minneapolis, MN: Pearson.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.
Kazdin, A. E. (1979). Situational specificity: The two-edged sword of behavioral assessment. *Journal of Psychopathology and Behavioral Assessment, 1*, 57–75.
Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child and Adolescent Psychology, 34*, 548–558.
Kenny, D. A., & Kashy, D. A. (1992). Analysis of multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.
Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
Langland, S., Lewis-Palmer, T., & Sugai, G. (1998). Teaching respect in the classroom: An instructional approach. *Journal of Behavioral Education, 8*, 245–262.
Marsh, H., & Bailey, M. (1991). Confirmatory factor analysis of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*, 47–70.
Marsh, H., Smith, I., & Barnes, J. (1983). Multitrait-multimethod analyses of the self-description questionnaire: Student-teacher agreement on multidimensional

ratings of student self-concept. *American Educational Research Journal, 20*, 333–357.
Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradovic, J., Riley, J. R., & Tellegen, A. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology, 41*, 733–746.
McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications* (3rd ed.). New York, NY: Springer.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.
Miller, F. G., Cohen, D., Chafouleas, S. M., Riley-Tillman, T. C., Welsh, M. E., & Fabiano, G. A. (2015). A comparison of measures to screen for social, emotional, and behavioral risk. *School Psychology Quarterly, 30*, 184–196.
Miller, F. G., Riley-Tillman, T. C., Chafouleas, S. M., & Schardt, A. A. (2017). Direct behavior rating instrumentation: Evaluating the impact of scale formats. *Assessment for Effective Intervention, 42*, 119–126.
Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171–189.
Patterson, G. R., Forgatch, M. S., Yoerger, K. L., & Stoolmiller, M. (1998). Variables that initiate and maintain an early-onset trajectory for juvenile offending. *Development and Psychopathology, 10*, 531–547.
Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children* (2nd ed.). Circle Pines, MN: AGS Publishing.
Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior assessment system for children* (3rd ed.). Minneapolis, MN: Pearson.
Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. M., & Glazer, A. D. (2008). Examining agreement between Direct Behavior Ratings (DBRs) and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions, 10*, 136–143.
Roberts, M. A., Milich, R., Loney, J., & Caputo, J. (1981). A multitrait-multimethod analysis of variance of teachers1 ratings of aggression, hyperactivity, and inattention. *Journal of Abnormal Child Psychology, 9*, 371–380.
Saylor, C. F., Finch, A. J. J., Furey, W., Baskin, C. H., & Kelly, M. M. (1984). Construct validity for measures of childhood depression: Application of multi-trait–multimethod methodology. *Journal of Consulting and Clinical Psychology, 52*, 977–985.
Shapiro, E. S. (2004). *Academic skills problems workbook* (Revised ed.). New York, NY: Guilford Press.
Shapiro, E. S. (2011). *Academic skills problems: Direct assessment and intervention.* New York, NY: Guilford Press.
Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools, 41*, 551–561.
Shapiro, E. S., & Kratochwill, T. R. (2000). *Conducting school-based assessments of child and adolescent behavior.* New York, NY: Guilford Press.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.
Silva, F. (1993). *Psychometric foundations and behavioral assessment.* Newbury Park, CA: Sage.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.
Spilt, J., Koomen, H., Stoel, R., Thijs, J., & van der Leij, A. (2011). Teachers' assessment of physical aggression with the preschool behavior questionnaire: A multitrait-multimethod evaluation of convergent and discriminant validity. *Journal of Psychoeducational Assessment, 29*, 407–417.
Sprick, R., Booher, M., & Garrison, M. (2009). *Behavioral response to intervention.* Eugene, OR: Pacific Northwest.
Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology, 22*, 107–116.
Stormshak, E. A., Bierman, K. L., McMahon, R. J., Lengua, L. J., & Conduct Problems Prevention Research Group (2000). Parenting practices and child disruptive behavior problems in early elementary school. *Journal of Clinical Child Psychology, 29*, 17–29.
Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Erlbaum.
Sugai, G., & Horner, R. (2002). The evolution of discipline practices: School-wide positive behavior supports. *Child and Family Behavior Therapy, 24*, 23–50.
Walker, B., Cheney, D., Stage, S., & Blum, C. (2005). Schoolwide screening and positive behavior supports: Identifying and supporting students at risk for school failure. *Journal of Positive Behavior Interventions, 7*, 194–204.
Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10*, 205–212.
Whitcomb, S. A., & Merrell, K. W. (2012). *Behavioral, social, and emotional assessment of children and adolescents* (4th ed.). New York, NY: Routledge.
Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26.
Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25*, 9–23.
Wolfe, V. V., Finch, A. J., Saylor, C. F., Blount, R. L., Pallmeyer, T. P., & Carek, D. J. (1987). Negative affectivity in children: A multitrait-multimethod investigation. *Journal of Consulting and Clinical Psychology, 55*, 245–250.