Examining Developmental Differences in Teachers' Observed Classroom Management Strategies Across Elementary, Middle, and High School

Heather L. McDaniel[1], Summer S. Braun[2], Jessika H. Bottiani[1], Danielle De Lucia[3], Patrick H. Tolan[1], and Catherine P. Bradshaw[1, 3]

[1]School of Education and Human Development, University of Virginia

[2]Department of Psychology and Center for Youth Development and Intervention, University of Alabama

[3]Bloomberg School of Public Health, Johns Hopkins University

McDaniel, H. L., Braun, S. S., Bottiani, J. H., De Lucia, D., Tolan, P. H., & Bradshaw, C. P. (2022). Examining Developmental Differences in Teachers' Observed Classroom Management Strategies Across Elementary, Middle, and High School. School Psychology Review, 1-21 https://doi.org/10.1080/2372966X.2022.2100275

Published in *School Psychology Review*.

Author Note

Heather L. McDaniel https://orcid.org/0000-0001-8214-9551; Summer S. Braun https://orcid.org/0000-0001-6203-4845; Jessika H. Bottiani https://orcid.org/0000-0001-7810-1707; Danielle De Lucia; Patrick H. Tolan https://orcid.org/0000-0001-5669-8442; Catherine P. Bradshaw https://orcid.org/0000-0003-2048-3225

Funding: The research reported here was supported by the Institute of Education Sciences under award numbers R324A110107 and R305A150221, the National Institute of Justice under award numbers 2014-CK-BX-005 and 2015-CK-BX-008, the U.S. Department of Education and Spencer Foundation, awarded to Catherine Bradshaw at the University of Virginia and Johns Hopkins University. Additional support was provided by the Robert Wood Johnson Foundation, awarded to Patrick Tolan at the University of Virginia. The opinions expressed are those of the authors and do not represent views of the funders.

Acknowledgements: The authors would also like to acknowledge the research team members and teachers that helped to make this work possible. We also thank Dr. Antonio Morgan-López (RTI International) for consultation on our statistical procedures.

All authors of this article declare they have no conflicts of interest. We have complied with ethical principles in our treatment of individuals participating in the research and the research has been approved by appropriate institutional review boards.

Correspondence regarding this article should be addressed to Heather L. McDaniel, School of Education and Human Development, University of Virginia, P.O. Box 400281, Charlottesville, VA 22904. Email: hm8tc@virginia.edu

Abstract

Classroom management practices are critical to the success of teachers and students, and a growing number of programs have been developed to improve these practices. However, there has been less investigation into observational tools to assess classroom management and explore whether it can be measured consistently by observers across elementary, middle, and high school classrooms. Moreover, there is a need to determine how classroom management practices vary as a function of school settings and classroom contexts (e.g., class size and racial composition). The current study aimed to examine classroom management practices using the Assessing School Settings: Interactions of Students and Teachers (ASSIST; Rusby, Taylor, & Milchak, 2001), an observational measure administered by trained external observers across 3,263 classrooms. A series of analyses indicated that the ASSIST demonstrated partial MI across contexts, and was particularly robust across class size and racial composition, which enabled us to contrast latent mean differences across developmental levels. Latent means of classroom practices across elementary and middle school were similar, whereas elementary school and high school classrooms differed significantly. The findings provide evidence that the ASSIST is similarly measuring classroom management across classroom contexts but is sensitive to mean differences in classroom management across classroom contexts.

Impact Statement: The Assessing School Settings: Interactions of Students and Teachers (ASSIST) is an observational measure of classroom management that is helpful for assessing differences in classroom management constructs across classroom levels (i.e., elementary, middle, high), as well as select contextual variables, including class size and racial composition. Findings illustrate the promise of the ASSIST as a tool for measuring teachers' classroom management practices and detecting differences across these settings, suggesting it may be particularly useful in large scale studies which span multiple grade and school levels.

Keywords: classroom management; student behavior; observation; classroom assessment; measurement invariance

Examining Developmental Differences in Teachers' Observed Classroom Management Strategies Across Elementary, Middle, and High School

The classroom is a critical ecological context for promoting positive student development (Bronfenbrenner, 1979; Bronfenbrenner & Morris, 2007; Eccles & Roeser, 2011). Cultivating a supportive classroom context requires teachers to effectively manage the social, emotional, and behavioral dynamics of the classroom. Effective classroom management has been conceptualized as a series of best practices consisting of foundational, prevention, and responsive classroom management components that support class-wide academic engagement and competence (Collier-Meek et al., 2019). Use of such classroom management strategies has been associated, in both correlational and experimental research, with students' social-emotional development, behavioral health, and academic achievement (e.g., Allen et al., 2013; Downer et al., 2012; Korpershoek et al., 2016; Mashburn et al., 2008; Oliver et al., 2011). Despite being critical to a teacher's mission, classroom management is an area in which teachers receive minimal formal pre-service training (Hirsch et al., 2021) and is reported to be a major source of stress (Clunies-Ross et al., 2008; Freeman et al., 2014). As such, there has been increased interest in programs, interventions, and policies to improve teachers' use of effective classroom management practices; however, there has been less attention to the creation of valid and reliable measures necessary to assess those processes in real time and across settings (e.g., elementary, middle, and high school classrooms).

Observational measures have emerged as an increasingly appealing option to reduce shortcomings associated with traditional student- and/or teacher-reported measures (Aguiar & Aguiar, 2020; Pianta & Hamre, 2009; What Works Clearinghouse, 2020). As such, an important and critical next step for the field is to empirically assess the extent to which core dimensions of effective classroom management can be conceptualized consistently by independent observers across classroom contexts. Specifically, there is a need for psychometrically sound classroom observational measures that can be efficiently used across multiple grade and school levels, thereby allowing for comparisons across contexts. Thus, identifying a single measurement tool that can be utilized across classroom contexts (i.e.,

class size, classroom racial composition) and school types (i.e., elementary, middle, high) is of practical importance for both researchers as well as practitioners. For example, such an observational measure would be helpful in large randomized controlled trials (RCTs) and district- or state-initiated program evaluations that include multiple schools, with varying classroom characteristics, as well as those spanning multiple grades and school levels. Moreover, such a tool could help identify teachers in need of professional development and coaching to improve classroom management practices across classroom contexts.

The current study aimed to address this measurement need by studying the Assessing School Settings: Interactions of Students and Teachers (ASSIST; Rusby et al., 2001, 2011), an observational measure of classroom management practices administered by trained external observers that has shown promise for use in large multi-grade and multi-school applications (e.g., Gaias et al., 2019; Larson et al., 2021; Pas et al., 2015; Tolan et al., 2020). We leveraged data from the ASSIST across multiple studies to determine whether teachers' classroom management strategies can be measured in the same way by independent observers across elementary, middle, and high school classrooms, as well as class size and classroom racial composition, by testing for measurement non-invariance/ differential item functioning (MNI/DIF). In addition, upon establishing measurement invariance (MI), we were substantively interested in potential developmental differences that may be detected using such a measure, as well as exploring how other classroom contextual variables (i.e., class size and racial composition) are associated with levels of ASSIST classroom management constructs.

Background on the ASSIST

Observational assessments of teacher and student classroom behavior by independent, trained observers have been increasingly used as an alternative or supplement to teacher self-report measures, given limitations associated with teacher self-report data (Aguiar & Aguiar, 2020; Debnam et al., 2015; Larson & Bradshaw, 2017; Pas et al., 2011; Pianta & Hamre, 2009). Observations are at times considered burdensome and expensive (Kane & Staiger, 2012), as such some researchers have opted for less costly alternatives (e.g., teacher ratings of students, teacher-reported logs of behavior; Rowan & Correnti, 2009).

However, standardized observations, as compared to teacher- and student-report of classroom practices, have demonstrated the strongest effect sizes on student outcomes in a meta-analysis covering a decade of teacher effectiveness research (Seidel & Shavelson, 2007) and reduce the potential bias introduced by teachers performing the observations. This is a particular concern in school-based studies where teachers are often delivering the intervention, and thus not unaware of the students' intervention status (What Works Clearinghouse, 2020). As such, observations have proven to be useful for research in widescale studies of classroom practices. Moreover, expert observation can be a particularly helpful coaching tool, allowing the observer to garner understanding of a teacher's strengths and weaknesses to set targets for performance feedback and coaching (Reinke et al., 2011, 2015).

A variety of observational tools have been created to assess classroom management that differ in method and focus. For example, some classroom management observation systems were initially created to focus on classroom management practices as they relate to a specific target student (e.g., Setting Factors Assessment Tool [Stichter et al., 2004, 2009]; Teacher-Pupil Observation Tool [T-POT; Martin et al., 2010]), although many of these measures have also been adapted or extended to focus more broadly on global teacher classroom management as well (e.g., Hickey et al., 2017). Others were initially developed to focus on the classroom as a whole (e.g., Classroom Assessment Scoring System [CLASS]; La Paro et al., 2004; Pianta et al., 2008). However, most classroom management observational measures have focused on classrooms of younger students (e.g., elementary school) while measures for secondary classrooms often rely on teacher- and student-report (Pianta & Hamre, 2009).

The ASSIST is one of only a few observational measures that has been used across elementary, middle, and high school classrooms (see discussion of exceptions below; e.g., CLASS [La Paro et al., 2004]). The ASSIST measure was originally developed by Rusby and colleagues (2001; 2011) to capture an independent perspective on the classroom, with a focus on teacher practices and student behaviors that reflect effective classroom management practices. As compared to many other observational tools (Praetorius & Charalambous, 2018), it was intended to be a low-inference assessment of both teacher practices and student behaviors. Specifically, the ASSIST consists of two types of behavioral assessments

of the classroom: *tallies* (i.e., running counts) of specific teacher and student behaviors and Likert-style *global ratings* of teacher and student behaviors. The tallies are collected in real time as a prime for the global ratings, which are completed immediately following the classroom observation.

The global ratings are organized into five theoretically-derived subscales that reflect the foundational, preventative, and responsive nature of effective classroom management practices (as described by Collier-Meek et al., 2019). Foundational classroom management practices create the basis for a well-functioning classroom, including the physical setup of the room, classroom routines, and positively stated expectations. The Teacher Direction and Influence scale of the ASSIST and the Teacher Proactive Behavior Support scale assess teachers' use of foundational practices. Teacher Direction and Influence assesses for evidence of classroom routines and how the teacher influences student behavior. The Proactive Behavior Support scale includes items related to provision of clear instructions and learning objectives, as well as teachers' use of labeled praise.

Preventive classroom management practices help teachers actively prevent problem behaviors from occurring, such as through active supervision, creating opportunities to respond, and offering precorrections. The ASSIST assesses preventative practices in three scales: Teacher Monitoring, Teacher and Student Meaningful Participation, and Teacher Anticipation and Responsiveness. Teacher Monitoring assesses whether a teacher positions themselves to engage in active supervision of student behavior. Teacher and Student Meaningful Participation assesses whether teachers provide students opportunities to respond, make choices, and take leadership roles, as well as prosocial classroom behavior amongst students. Teacher Anticipation and Responsiveness focuses on teacher precorrections and anticipation of student behavioral and academic difficulties. In addition, responsive classroom management practices reinforce appropriate behavior and discourage problem behavior. The Teacher Anticipation and Responsiveness scale of the ASISST spans both preventative and responsive practices by additionally assessing teacher responsiveness to student needs.

Finally, the global ratings of student classroom behavior are organized into two subscales: Student Cooperation and Student Socially Disruptive Behavior. Student Cooperation assesses whether

students comply with rules, are respectful, and demonstrate academic readiness. Student Socially Disruptive Behavior focuses on a range of disruptive behaviors in the classroom including social conversations occurring when they should not be, irritability and sarcasm, arguing, verbal aggression, and bullying. For additional details on the ASSIST observational measure, see the Methods section.

The ASSIST has demonstrated evidence of validity in previous applications of the measure. As would be expected based on the ASSIST's hypothesized nomological net (Cronbach & Meehl, 1955), prior research exploring how teachers' classroom management practices, as measured by the ASSIST, are associated with student behaviors has indicated that greater teacher influence, responsiveness, monitoring, proactive behavior support, and meaningful engagement are associated with increased student cooperation and decreased disruptive behavior (e.g., Gaias et al., 2019; Larson et al., 2018, 2021; Pas et al., 2015). In addition, the ASSIST has been used to assess the impact of interventions focused on promoting a supportive classroom setting, and has demonstrated sensitivity to changes in teacher and student behaviors (Bradshaw et al., 2018; Tolan et al., 2020). While the ASSIST has demonstrated strong psychometric properties across many contexts and applications (Bradshaw et al., 2018; Debnam et al., 2015; Gaias et al., 2019; Pas et al., 2015; Tolan et al., 2020), there has yet to be a systematic investigation of the MNI/DIF of the ASSIST; moreover, a test of latent mean differences across elementary, middle, and high school classrooms or other classroom contexts is also a necessary step to inform how the ASSIST may be utilized across multiple, diverse classroom contexts.

Assessing Measurement Invariance and Testing Latent Mean Differences

Contemporary conceptualizations of classroom management and related classroom constructs, as measured by observational systems like the ASSIST, often involve organizing teacher and student classroom behaviors as observed indicators of theoretically-informed latent factors (Hamre & Pianta, 2007). To compare latent means across groups, such as across types of classrooms, measurement invariance, or *at least partial invariance*, must be established to ensure that the construct is operationalized in the same or a similar manner across groups, making subsequent comparisons of latent means across groups meaningful (Borsboom, 2006; Byrne et al., 1989; Putnick & Bornstein, 2016).

Partial MI is demonstrated when some but not all parameters in the measurement model are constant across groups/settings (additional detail provided below; Byrne et al., 1989; Vandenberg & Lance, 2000), allowing for a similar but not identical construct parameterization across groups/settings.

Districts and states engaged in scale-up initiatives would benefit from the availability of a measure that could be consistently utilized districtwide or statewide across classroom contexts, from elementary through high school. There are many large-scale initiatives to scale-up tiered frameworks that focus on promoting student success in schools (e.g., national efforts to disseminate Positive Behavior Interventions and Supports [PBIS]). Many of these frameworks and embedded interventions for promoting a positive school climate and student social, emotional, behavioral, and academic outcomes include a classroom behavior management component (e.g., Barrett et al., 2008; Bradshaw & Pas, 2011; Horner et al., 2014). For example, the state of Maryland adopted PBIS, a three-tiered approach for prevention and promotion of positive behavior in schools, as their statewide K-12 response to intervention approach for behavior (Barrett et al., 2008; Bradshaw et al., 2012). Within the PBIS Maryland model, the universal tier 1 supports include classroom management systems and practices (Barrett et al., 2008); the statewide partnership conducted evaluations of PBIS spanning elementary (e.g., Bradshaw et al., 2010) through high schools (e.g., Bradshaw et al., 2021). In this context, it would have been exceptionally helpful to use a consistent measurement approach, like the ASSIST, across all settings to compare potential differential impacts of PBIS across school levels (see Pas et al., 2019).

Moreover, there is growing interest in contrasting the impact of various classroom management programs to determine what works for whom and under what conditions. However, such comparisons are not appropriate or meaningful unless the measure has first demonstrated (at least partial) measurement invariance, meaning that the measurement of the latent construct does not vary substantially based on school level or other classroom characteristics (e.g., class size, racial composition). Moreover, it would likely be most efficient to be able to compare scores on the same measure across classroom contexts, to target teacher training or coaching efforts to particular classroom contexts or to evaluate in which contexts the initiative is yielding the largest impacts. Yet to date, there have been no such studies which

have explored measurement invariance of classroom management practices across such a wide range of classroom characteristics and school levels.

To address this gap, one must first provide evidence of measurement invariance of a classroom management assessment tool with regard to classroom context (e.g., for an overview of MI see Bauer, 2017; Kline, 2015). At the weakest level, configural invariance is established when the structure of the measurement model is consistent across groups, but all parameters are freely estimated. When configural invariance holds, it suggests that there are the same number of latent constructs across groups and that the mapping of items to latent factors is the same. Next, metric invariance builds on configural invariance and is established when factor loadings are statistically equivalent across groups. Metric invariance suggests that the items are weighted in the same way across groups and that factor scores could be calculated using the same factor loadings. MNI/DIF may be present when items are more or less representative of the latent construct across the groups being compared. Although full MI is preferable and indicates statistically consistent measurement properties across contexts, there is an expectation of at least partial MI to allow for contrasts across settings. Specifically, partial MI refers to situations where some but not all of the loadings are non-invariant across groups (Byrne et al., 1989; Vandenberg & Lance, 2000). In this case the items demonstrating DIF are freely estimated across groups, accounting for cross-group differences in measurement model parameters. The next level of MI, scalar invariance, or strong invariance, holds when both item loadings and item intercepts are equal across groups. Scalar invariance suggests that item means are statistically equivalent across groups, controlling for the level of the latent factor. Partial MI also holds when some but not all item intercepts are equal across groups. In this case, non-invariant parameters are freely estimated across groups.

Establishing complete or partial MI across classroom contexts would suggest that classroom management constructs are manifested similarly enough across those contexts and that the measure can be used across those contexts to make meaningful comparisons (Byrne et al., 1989; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Here partial MI is demonstrated when some but not all parameters (i.e., loadings, intercepts) in the measurement model are constant across groups/settings (Byrne et al.,

1989; Vandenberg & Lance, 2000). Importantly, establishing measurement invariance or partial invariance has important implications for use of the measure. When a measure is invariant across classroom developmental levels, it suggests that the user can score teachers in elementary school classrooms in the same manner that they would score teachers in a middle or high school classroom. However, when there is DIF attributable to school level, the scoring algorithm would need to accommodate for differences in measurement across developmental levels, to appropriately assess a teacher's level of the latent construct. Once scores account for this DIF, partial invariance allows latent means to be compared across groups with a meaningful interpretation (Byrne et al., 1989; Vandenberg & Lance, 2000).

Developmental Differences in Teachers' Classroom Management Strategies

Observational Measures

Although many measures of classroom management focus on elementary and middle school, one of the most commonly used and well-validated classroom observational tools, the Classroom Assessment Scoring System (CLASS; La Paro et al., 2004; Pianta et al., 2008), has been validated and widely-utilized in pre-school settings through secondary schools. The CLASS measures the quality of teachers' interactions with students in the classroom, a construct closely related to classroom management, and is traditionally organized into three domains: emotional support, classroom organization, and instructional support. The classroom organization domain is comprised of subdomains aligned with classroom management (i.e., behavior management, productivity, instructional learning formats; La Paro et al., 2004; Pianta et al., 2008; Pianta & Hamre, 2009). These domains are thought to be supportive for students across school levels; however, the CLASS has specific versions of the measure, tailored to the developmental level of the students (Infant, Toddler, Pre-Kindergarten, Kindergarten – 3rd Grade, Upper Elementary, and Secondary; Teachstone, 2011). These different versions of the CLASS were developed based on theory to be developmentally appropriate (Allen et al., 2013; Teachstone, 2011). While the CLASS's developmental focus is advantageous on many levels, the differences across the various versions preclude examination of MI across elementary, middle, and high school classrooms on the

CLASS. Thus, although the core underlying structure of the CLASS is generally consistent in kindergarten through 12th grade (Allen et al., 2013; Teachstone, 2011), suggesting that measurement of classroom management may also be similar across school levels, we are unable to directly compare the results of tests of MI on the CLASS across school levels.

Student- and Teacher-Report

Given the relative paucity of observational studies of classroom management which use the same measure and span elementary through high school settings, we also consider possible developmental differences based on student- and teacher-reported data from related measures of classroom practices, climate, and relationships. For example, a series of studies of German students suggested that teachers' classroom practices, as rated by students, are similar from late elementary through secondary classrooms (Gaertner & Brunner, 2018; Wisniewski et al., 2020). Both of these studies provided evidence of MI prior to examining the developmental differences. Specifically, in the first of these studies, which included domains reflective of teacher influence, proactive behavior support, and responsiveness, MI was demonstrated from grade three through high school (Gaertner & Brunner, 2018). In the second study, which included domains related to teacher influence and proactive behavior support, as well as student engagement, MI was supported across grades 5-12 (Wisniewski et al., 2020). Moreover, other self-report measures of related constructs, such as student engagement and school climate have demonstrated invariance across grade levels (Betts et al., 2010; Waasdorp et al., 2020), and demonstrated some variation in latent means by school level (Waasdorp et al., 2020).

Yet other studies of teacher-reported indices have documented MNI/DIF, such as the conflict subscale of the Student-Teacher Relationship Scale (STRS; Pianta, 2001). As one might expect, certain developmentally sensitive items, such as "this child whines or cries when he/she wants something" loaded more strongly for younger students than older students, whereas other items, like "this child feels that I treat him/her unfairly" loaded more strongly for older students than younger students (Koomen et al., 2012), indicating that some items were a more salient reflection of the construct in one age group than the other. Similar results were found for the closeness subscale of the STRS measure. Additionally, the

intercepts of some items differed by student age (e.g., “this child sees me as a source of punishment and criticism” was higher for older students than younger students), suggesting that across ages, controlling for the level of the latent construct, item means were not equivalent.

Mean Differences by Developmental Level

In addition to the need for assessment of MI, latent means of teacher practices and student behaviors may differ across classroom developmental levels. Previous research has found differences in mean levels of teacher practice and student behavior across elementary school: classes with younger students had higher rates of classroom rule violations (i.e., disruptive behavior) than classes of older students, and teachers of younger students demonstrated higher rates of praise (i.e., approvals) than teachers of older students but teachers demonstrated similar rates of effective command usage across grades (Owens et al., 2018). Other research across elementary and middle school suggests that middle school teachers provide greater “instructional management” (i.e., monitoring, structuring routines, facilitating participatory approaches to instruction) but there were no significant differences in the “behavior management” latent means (i.e., establishing rules and a reward system, providing opportunities for student input) across elementary and middle school classrooms (Sass, 2011).

This comparison of classroom management mean differences across elementary, middle, and high school may provide insight into strengths and potential intervention targets at each developmental level. As such, this research has important implications for the development and assessment of evidence-based interventions aimed at improving teachers’ use of classroom management strategies that promote socially, emotionally, behaviorally, and academically supportive classroom environments. Given no identified research that compares mean levels of teacher and student behaviors across elementary, middle, and high school classrooms, additional research exploring the extent to which aspects of teacher classroom management vary developmentally is needed. The ASSIST holds promise as an observational measure for providing insight regarding these research gaps related to developmental differences in classroom management. Further, should differences exist, this may highlight the need for additional consideration of

developmental differences in tailoring professional development for teachers related to classroom management (for a review see Oliver et al., 2011; Simonsen et al., 2008).

Exploring Broader Classroom Contextual Factors

In addition to examining developmental differences across school levels, we were also interested in whether MI varied as a function of select classroom contextual factors. Indeed, the measurement and levels of teachers' classroom management practices may also vary as a function of contextual factors in the classroom beyond school level (Downer et al., 2010). Although the main focus of this study largely centered on how developmental level may be related to measurement invariance of the ASSIST, in addressing this research question we also explored whether class size and classroom racial composition were associated with measurement invariance and subsequently, levels of latent classroom management practices.

Class Size

There has been great interest in class size, which relates to student-teacher ratio, as a possible factor contributing to levels of classroom management, with the general expectation that smaller classes are easier to manage. For example, the Tennessee Class Size Project and others determined that students excelled academically when classes were smaller (Finn & Achilles, 1990, 1999; Mosteller, 1995). By virtue of the number of students in the class and teachers' finite attentional, cognitive, and behavioral resources, it is reasonable to question how class size may be related to the measurement and level of teachers' classroom management practices.

Class Size and MI. An investigation into the processes underlying the positive effects of small class size on student achievement led to an examination of whether teachers' classroom management practices and interactions with students may vary by class size (Bourke, 1986; Evertson & Randolph, 1989; Folmer-Annevelink et al., 2010; Hargreaves et al., 1998). Despite the use of observational measures (e.g., Teacher Record) in classes of varying sizes to answer this question, we know of no studies to have documented MI work with these measures. Indeed, even outside of this specific line of investigation, MI of observational measures of teachers' classroom practices based on class size appears exceedingly rare, a

noted limitations of cross-context educational research (Kalender, 2015; Oude Groote Beverborg et al., 2021). Some of the most relevant work regarding MI of school-based measures by group size has been done with regard to MI by school size in a series of studies on measures of student-reported temptations for using tobacco and alcohol; these studies found evidence of strong MI across schools of small and larger sizes (i.e., less than 200 students versus greater than 200 students) (Babbin et al., 2011; Harrington et al., 2011; McGee et al., 2012). The relative paucity of research in this area led us to explore whether class size was related to MI but offered no directional hypotheses.

Mean Differences by Class Size. Despite no known measurement work on the observational systems used in the subsequently discussed studies, researchers continued to address how teachers' practices and interactions with students may differ by class size. This body of work has shown that teachers' use of routine classroom management practices tend to be higher in classes of larger sizes, whereas teachers in smaller classes spent more time on task-related content and inquiry (Folmer-Annevelink et al., 2010; Hargreaves et al., 1998). Correspondingly, it has also been noted that students' engagement is greatest when class sizes are smaller (Blatchford et al., 2011). Thus, we may observe greater mean levels of teachers' classroom management practices in larger classroom contexts, and higher student engagement in smaller classes.

Classroom Racial Composition

There is growing interest to ensure classroom management practices are culturally responsive to all learners (Gaias et al., 2019; Larson et al., 2018). A necessary precursor to examining these relationships is exploring whether classroom management practices are assessed similarly across classrooms with differing racial compositions.

Classroom Racial Composition and MI. To our knowledge, few studies have explored the role of MNI/DIF as a factor potentially underlying differences in *observed* teachers' use of classroom management practices based on the visible racial composition of the classroom. In one related study examining classroom racial composition, specifically in the proportion of Latinx students in a classroom, Downer and colleagues (2012) found that the CLASS demonstrated strong MI across more than 700 pre-k

classrooms in 11 states. This preliminarily suggests that classroom management practices may be measured equivalently across classroom racial compositions but warrants further investigation.

Mean Differences by Classroom Racial Composition. In terms of levels of classroom management practices, some research has found that teachers with a higher percentage of White students in their classrooms were more likely to exhibit high-quality classroom management (Gaias et al., 2019). In the latter study, students were ascribed race based on their visible features by an external observer, consistent with a critical race theory perspective (Solorzano, 1997), which suggests that teachers may privilege White youth based on their skin color and other phenotypic traits. As such, teacher classroom management could vary as a function of the visible *racial composition* of students in the classroom. Specifically, some research has found that bias in teacher perceptions of students' behavior can advantage White students (compared to students of color) (Downey & Pribesh, 2004). Other studies have similarly found that teachers' observed use of effective classroom management practices is higher in classrooms with a greater proportion of students perceived to be White (Authors, under review; Bottiani et al., 2019). However, without assessing measurement invariance, the source of these differences may be attributable to measurement differences, rather than differences in levels of the latent constructs.

Current Study

Although observations are frequently considered a "gold standard" for assessing the impacts of school-based programs, there is little MI evidence to support the use of a particular observational tool to measure teachers' classroom management practices across elementary, middle, and high school classrooms. Our first research aim was to examine MI of the ASSIST across elementary, middle, and high school classrooms. We were largely confirmatory in our MI hypotheses for the first aim given prior studies suggesting that student-reported classroom practices and related constructs are invariant across elementary, middle, and high school classrooms (Betts et al., 2010; Gaertner & Brunner, 2018; Waasdorp et al., 2020; Wisniewski et al., 2020). However, it is possible that certain aspects or dimensions of teacher-student classroom interactions may be influenced by the students' developmental capabilities, thereby suggesting MNI/DIF (Allen et al., 2013; Teachstone, 2011; Watts-Taffe et al., 2012). For our

second aim, we examined how teacher and student classroom behaviors, as evidenced by latent factor means, varied across these settings. We hypothesized that elementary school teachers would demonstrate more frequent use of active classroom management strategies, compared to secondary school classrooms. Under our third aim, we explored MI and latent mean differences related to additional classroom contexts, class size and racial composition. Given the paucity of prior empirical work in this area and the exploratory nature of this aim, we did not have specific hypotheses.

Method

Sample

Data for the current paper were drawn from baseline assessments across seven research projects. The analytic sample consisted of data from 3,263 classrooms nested within 185 schools. Twenty-six of the cases from the original sample were listwise deleted (i.e., 0.8% of the original sample), as they were missing predictor information. The participating schools were located in two states in the Southern Region of the United States and ranged in enrollment from 188 to 2,285 students, with an average of 814 students enrolled per school. Students enrolled in participating schools were, on average 41.61% African American (*SD* = 27.04%), 11.58% Latinx (*SD* = 11.88%) and 37.96% White (*SD* = 26.10%). And, on average, 51.06% (*SD* = 22.91%) of students in a school received free or reduced priced meals. Of the participating classroom teachers, 18.88% of teachers (*n* = 616) were in elementary school classrooms, 38.49% in middle school classrooms (*n* = 1,256), and 42.63% in high school classrooms (*n* = 1,391). Teachers were predominately female and, on average, were in classrooms with 21 students, of which 41.75% were White and 50.06% male (see Table 1 for additional information about participating teachers and classrooms).

Procedure

This study leveraged classroom observational data from seven RCTs of school-based interventions (see Bradshaw et al., 2014, 2018; Pas et al., 2019) in an integrative data analysis (Curran & Hussong, 2009). Although the data were from separate projects, the same training procedures and administration of observations were coordinated by a shared research team. Trial participation involved

permitting a trained ASSIST observer from the research team to conduct classroom observations. To avoid the influence of intervention effects, only data from baseline assessments were used in the present study. Participation by schools and teachers in the trial was voluntary. The researchers' Institutional Review Boards approved each of the RCTs.

ASSIST Training

The same training procedures were used across all studies from which the ASSIST data were drawn. All observers were trained to reliability and certified in the ASSIST. Independent observers were contracted and completed an 8-hour didactic training, including extended video coding practice and feedback cycles. After the didactic training, observers underwent in-school live practice with an expert ASSIST observer. Observers then completed a live, in-school reliability assessment. Observer trainees were deemed certified and reliable in the ASSIST after passing the in-school reliability assessment with 80% or higher inter-rater reliability with expert observers across student and teacher tallies. Midway through the baseline data collection for the trials, observers completed field-based or video-based reliability assessments to recalibrate and correct any rater drift. At midpoint recalibration, observers were allowed to continue observations only after again meeting or exceeding 80% reliability with a master observer. Prior research on the ASSIST indicated that when a teacher was administered multiple observations, intraclass correlations suggested little variability across administrations. Specifically, intraclass correlations of the ASSIST scales ranged from 0.72 to 0.81, with an average of 0.75, suggesting relatively little variability across three observations and that most of the variance was nested at the teacher-level (see Gaias et al., 2019). For additional details on the training and reliability of the observers, see Bradshaw et al. (2018), Debnam et al. (2015), Gaias et al. (2019) and Pas et al. (2015).

Administration of the ASSIST

Data were collected using a handheld Samsung tablet and the Pendragon Forms Software Program. Per the ASSIST administration protocol (Rusby et al., 2001, 2011), upon entering the classroom, observers spent three minutes adjusting to the physical and social environment and recording descriptive information about the classroom. The observer then commenced with the 15-minute

observation window during which they kept a live tally of teacher and student behaviors. After conducting the tallies, the observers then stepped outside of the classroom and completed the global ratings. Due to logistical considerations, the timing of observations ranged from morning to afternoon across the seven projects. For additional details on the ASSIST administration procedures, see Bradshaw et al. (2018); Debnam et al. (2015); Gaias et al. (2019); and Pas et al. (2015).

Measures

ASSIST Measure

As noted above, the ASSIST is an observational measure of teachers' classroom management strategies that consists of both classroom behavioral tallies (i.e., frequency counts of specific teacher and student behaviors) and global ratings of classroom social processes (Rusby et al., 2001; 2011). The global ratings are 42 Likert-style items that are the focus of the current research. Items were designed to be theoretically organized into seven subscales focused on teachers' classroom management practices and student classroom behavior (see items in Tables 2 & 3). The items for each of the scales (with the exception of the Student Socially Disruptive Behavior items) were rated on a five-point Likert scale (i.e., 0 = "Never"; 1 = "Seldom"; 2 = "Some of the time"; 3 = "A lot of the time"; 4 = "Almost continuously").

Teacher Direction and Influence. The five-item influence subscale assesses the establishment of classroom routines and how well the teacher influences student behavior (Cronbach's $\alpha = .93$; Alpha calculated based on results described below in which Teacher Direction and Influence and Cooperation were combined).

Teacher Proactive Behavior Support. The four-item proactive behavior support subscale reflects how often teachers proactively utilize strategies that prevent and/or supportively redirect disruptive behavior (Cronbach's $\alpha = .73$).

Teacher Monitoring. The four-item monitoring subscale assesses the extent to which the teacher uses strategies such as positioning and scanning to monitor student behavior (Cronbach's $\alpha = .88$).

Teacher and Student Meaningful Participation. The nine-item meaningful engagement subscale captures student and teacher classroom behaviors, assessing teacher facilitation and scaffolding

of student involvement in the classroom as well as student behaviors that demonstrate their engagement (Cronbach's α = .87).

Teacher Anticipation and Responsiveness. The six-item responsiveness subscale assesses how frequently the teacher anticipates and is responsive to various student needs (Cronbach's α = .84).

Student Cooperation. The seven-item cooperation subscale assesses how often students comply with classroom expectations, are respectful, and cooperative (Cronbach's α = .93; Alpha calculated based on results described below in which Teacher Direction and Influence and Cooperation were combined).

Student Socially Disruptive Behavior. The seven-item disruptive behavior subscale captures a range of student socially disruptive behaviors from social conversations (i.e., occurring when they should not) to verbal aggression towards the teacher and peer bullying. Items were rated on a five-point Likert scale (i.e., 0 = "Rarely occurred (0 times)"; 1 = "Rarely occurred (1 time)"; 2 = "Occurred a few times (2-3 times)"; 3 = "Sometimes occurred (4-6 times)"; and 4 = "Often occurred (6+ times)") (Cronbach's α = .64).

Demographic Predictors

Predictors of ASSIST MNI/DIF and latent scores were classroom level (i.e., elementary, middle, and high school classrooms), number of students in the classroom, and classroom racial composition. Classroom level was dummy coded such that elementary school classrooms were the reference group. Number of students in the classroom was a count of students in the classroom by the ASSIST observer.

Classroom racial composition was also derived from ASSIST observations as the perceived number of White students in the classroom. For analyses, this count was divided by the total number of students in the classroom and multiplied by 100 to arrive at the percentage of students perceived to be White (i.e., possible range: 0 – 100%). The percentage of students perceived White was then divided by 10, such that a one unit increase on the scale represents a 10% increase in the percentage of students perceived to be White in the classroom. Note that observers were carefully trained prior to data collection regarding the distinction and use of this variable as *perceived* race (i.e., as differentiated from self-

identified race). Both the number of students and classroom composition variables were also mean centered for use in analyses.

Analytic Plan

Preliminary Analyses

Data preparation was conducted in IBM SPSS Statistics and R. In projects where there were multiple baseline assessments (e.g., some but not all projects included three baseline assessments per teacher), we utilized the sample_n function in the dplyr package (Wickham et al., 2020) in R to randomly sample one observation per teacher at baseline. Items were screened and removed from further analyses if multicollinearity diagnostics revealed a variance inflation factor (VIF) greater than ten (Kline, 2015). We conducted a descriptive assessment of the frequency of categorical responses across elementary, middle, or high school classrooms. In the case where one or more groups did not utilize a response option, and the other(s) did, response options were combined for all groups.

Measurement Invariance

Initial factor analysis was conducted to assess dimensionality in M*plus* (Version 8; Muthén & Muthen, 1998-2021), retaining all 3,263 classroom observations, using a categorical estimator to account for the ordered, categorical nature of the items (i.e., weighted least squares, mean and variance adjusted; WLSMV) and cluster-robust standard errors to account for nesting of classrooms within schools (i.e., using a sandwich estimator; Muthén & Satorra, 1995). We aimed to retain a simple structure in which each item mapped onto one respective subscale (i.e., Teacher Direction and Influence, Teacher Proactive Behavior Support, Teacher Monitoring, Teacher and Student Meaningful Participation, Teacher Anticipation and Responsiveness, Student Cooperation, and Student Socially Disruptive Behavior), as conceptualized in previous uses of this measure (Bradshaw et al., 2018; Pas et al., 2015; Rusby et al. 2011). Model fit was assessed using standard indices suggested in the methodological literature to indicate adequate fit of the model to the data when utilizing the WLSMV estimator (i.e., CFI $\geq$ .95, RMSEA $\leq$ .05) (Yu & Muthen, 2002).

After obtaining a configural model with adequate model-data fit, we utilized moderated nonlinear factor analysis (MNLFA) to assess MNI/DIF for each factor separately. MNLFA is a flexible measurement model that is an extension of structural equation modeling (Bauer, 2017; Bauer & Hussong, 2009) and allows examination of MNI/DIF (i.e., loading and intercept DIF) across multiple categorical and/or continuous predictors of MNI/DIF (for additional details about MNLFA see Bauer, 2017). We utilized maximum likelihood estimation, declaring ASSIST items as categorical, and cluster robust standard errors to account for the nesting of classrooms within schools (Muthén & Satorra, 1995). First, we automated a series of MNLFA models in MplusAutomation (Hallquist & Wiley, 2018), examining MNI/DIF for each item, with a separate model for each predictor and item combination. In each iteration of the model for a given factor, we estimated the same model with each item loading onto one factor as denoted by the final configural model, however, the predictor variable (i.e., developmental level, class size, classroom racial composition) and the focal item for which MNI/DIF was modeled, varied from model to model. The focal predictor variable in a given model predicted the loading and intercept of the focal item (i.e., capturing MNI/DIF), as well as levels of the latent factor (i.e., ASSIST scale). Factor loading MNI/DIF suggests that the items are not weighted in the same way across groups or levels of the predictor and that factor scores cannot be calculated using the same factor loadings across groups or levels of a predictor (Kline, 2015). Loading MNI/DIF may be present when items are more or less representative of the latent construct across the groups or levels of the predictor. Intercept MNI/DIF suggests that item intercepts are statistically different across groups, controlling for the level of the latent factor (Kline, 2015). When an MNI/DIF parameter was statistically significant at $p \leq .001$, the parameter was retained and specified in the penultimate global MNLFA model for each factor. Additionally, in the penultimate model, each covariate was specified to predict the mean level of the factor and project dummy codes accounted for nesting within project. A final MNLFA model was then run for each factor in which MNI/DIF parameters not statistically significant at the $p \leq .001$ in the penultimate model were pruned. The $p \leq .001$ statistical significance level was selected given the large sample size of teachers/classrooms.

Predicting Levels of the Latent Factors

Even in the presence of MNI/DIF, partial measurement invariance allows comparison of latent factor scores across groups and levels of continuous predictors, as long as the MNI/DIF is accounted for (Bauer, 2017). In the final global models for each factor, we modeled all MNI/DIF parameters that were identified as statistically significant at the $p \leq .001$ level in the individual item by predictor models. Additionally, each predictor was regressed on the latent factor (i.e., ASSIST scale) to ascertain how classroom contexts were associated with ASSIST scores, while accounting for partial MI.

Results

Preliminary Analyses

VIF

Assessment of the VIF indicated that there were no noteworthy multicollinearity issues (i.e., VIF > 10; Kline, 2015).

Consistency in Response Options

Due to sparseness of data (i.e., a given response option not being endorsed in one or more groups [elementary, middle, or high school classrooms]), five items had response scales reduced from a five-point to a four-point scale and one item had the response scale reduced from a five-point to a three-point scale (see valid ranges in Tables 2 & 3; items with a range less than zero to four indicate concatenated items).

Overview of the Model

The factor structure used for the current investigation was consistent with the theorized structure with two exceptions. Preliminary analyses in the current data indicated that the Influence subscale and Cooperation subscale correlated at a value greater than .90, suggesting that these subscales were not differentiating as distinct constructs. In subsequent invariance testing, Influence and Cooperation items were modeled as loading onto a single, combined construct. Second, through examination of modification indices, it was noted that one item, “Students respond to teacher’s questions and/or volunteer when

asked", demonstrated strong cross-loadings and perturbed model fit. This item was omitted from subsequent analyses.

MNLFA

First the configural model was fit to the data, in which the same factor structure was implied across all cases (i.e., all school levels, class sizes, and classroom racial compositions). Fit indices for the configural model suggested adequate model-data fit, χ^2 (764) = 4910.68, $p < .001$; *CFI* = .941; *RMSEA* = .041, suggesting that the model adequately fit the full sample data and implying configural MI. Next, MNLFA models were run for each ASSIST scale with results of the global MNLFA models presented in Table 4. Overall, eight items demonstrated statistically significant intercept MNI/DIF and five demonstrated statistically significant loading MNI/DIF (i.e., $p < .001$) in the penultimate model and were retained in the final MNLFA models. Other predictors of MNI/DIF that were statistically significant in individual predictor by item MNLFA models ($n_{items} = 20$), were non-significant in the penultimate MNLFA models and not retained in the final models.

School Level

The high school predictor (i.e., elementary school as reference) was most frequently associated with statistically significant DIF, predicting statistically significant intercept DIF for five items and loading DIF for two items. On the Influence and Cooperation scale, high school classrooms, as compared to elementary classrooms demonstrated a significantly higher than average intercept for the item, "Teacher has little/no control of or influence on students". High school classrooms also evidenced intercept DIF for three items on the Meaningful Engagement scale, with high school classrooms demonstrating significantly larger than average intercepts for the items, "Students praise and compliment one another" and "Students share their ideas/opinions" and a smaller than average intercept for the item, "Teacher gets students involved in lesson by asking questions or making comments" as compared to elementary classrooms. High school classrooms also demonstrated a significantly lower intercept value for the item, "Teacher maintains proximity to students who display a need for assistance or support" on the Responsiveness scale. In addition, high school classrooms reported smaller than average loadings for

the items, “The teacher clearly explains learning objectives prior to and/or during the lesson through summary or re-orientation statements” on the Behavior Support scale and “Teacher encourages students to share their ideas and opinions” on the Meaningful Engagement scale.

The middle school classroom indicator variable, as compared to elementary school classrooms, was also a statistically significant predictor of loading DIF for three items. On the Disruptive Behavior scale, there was a significantly greater than average loading in middle school classrooms, as compared to elementary school classrooms, for the item, “Social conversations occur between students and peers”. Middle school classrooms also demonstrated significantly greater loadings for the items, “Teachers is able to focus on one or two students while still scanning all other areas” on the Monitoring scale and “Teacher maintains proximity to students who display a need for assistance or support” on the Responsiveness scale.

Class Size

Class size was associated with intercept DIF for one item and loading DIF for one item on the Disruptive Behavior scale. Larger classrooms demonstrated a significantly larger intercept value for the item, “Social conversations occur between students and peers”. Larger classrooms also demonstrated a significantly greater loading for the item “Students are irritable or sarcastic toward peers”.

Classroom Racial Composition

Classroom racial composition was significantly predictive of intercept DIF for two items on the Responsiveness scale. There were higher-than-average intercepts for the items, “Teacher anticipates when students may have problems academically” and “Teacher notices when students have difficulty understanding a concept” in classrooms with a larger proportion of White students.

Predicting Levels of the ASSIST Latent Factors

The associations between classroom context variables and each ASSIST scale are presented in Table 5. We also present standardized effect sizes (i.e., small effect ≥ .2, medium effect ≥ .5, large effect ≥ .8).

School Level

Middle school, as compared to elementary school, classrooms did not demonstrate significantly different levels of scale factor means (i.e., at the $p \leq .001$ level). However, high school classrooms, as compared to elementary school demonstrated significantly lower levels of Behavior Support ($b = -.90$, $p < .001$), Monitoring ($b = -.64$, $p < .001$), Responsiveness ($b = -.89$, $p < .001$), and significantly higher levels of Disruptive Behavior ($b = -.64$, $p = .001$), with each of these effects ranging from moderate to large in magnitude (i.e., > .5).

Class Size

A greater number of students in the classroom was associated with lower levels of Influence and Cooperation ($b = -.02$, $p < .001$), Behavior Support ($b = -.02$, $p = .001$), Monitoring ($b = -.02$, $p < .001$) and Responsiveness ($b = -.03$, $p < .001$); however, each of these effects was negligible in magnitude (i.e., < .2).

Classroom Racial Composition

Classrooms with a greater proportion of White students demonstrated significantly greater Influence and Cooperation ($b = .06$, $p < .001$) and lower levels of disruptive behavior ($b = -.10$, $p < .001$); however, each of these effects was negligible in magnitude (i.e., < .2).

Discussion

Classroom management is a key aspect of a teacher's professional role. Moreover, classroom practices are often a key component of districtwide and statewide initiatives scale-ups of frameworks aiming to improve student social, emotional, behavioral, and academic outcomes (e.g., Positive Behavior Interventions and Supports; Bradshaw & Pas, 2011; Horner et al., 2014). As such, it is of both practical and theoretical relevance to understand how different classroom contexts may impact measurement properties of observational measures of classroom management. Although the ASSIST has demonstrated adequate psychometric properties in prior studies (Bradshaw et al., 2018; Debnam et al., 2015; Gaias et al., 2019; Pas et al., 2015; Tolan et al., 2020), these results have been based on data from samples largely comprised of high school and middle school classrooms; moreover, they have not considered how classroom contextual factors are associated with MNI/DIF. Thus, the current study explored whether the

ASSIST demonstrated MI across elementary, middle, and high school classrooms. Additionally, we aimed to understand how the levels of teachers' strategy use and students' classroom behavior varied across these settings. Given the dearth of existing research on this topic, the current study provides critical insight into which aspects of teacher classroom management vary across child and adolescent development and across classroom contexts. Having additional information regarding MI of the ASSIST would suggest it is appropriate to use to draw comparisons across school levels and context.

School Level

The results indicated that the ASSIST items demonstrated MNI/DIF based on school level, with the greatest frequency of statistically significant MNI/DIF attributed to high school, as compared to elementary school, and then middle school, as compared elementary school. Conceptually, these results suggest that the foundational, preventative, and responsive practices that teachers employ to effectively manage their classrooms (Collier-Meek et al., 2019), are disparately measured by the ASSIST across elementary, middle, and high school classrooms. This finding is inconsistent with research on other measures used across grade- and school-levels, such as a student-report measures of teacher classroom practices (Gaertner & Brunner, 2018; Wisniewski et al., 2020). However, it is consistent with measurement research on teacher-reported assessments of classroom management, which have reported both metric and scalar non-invariance across youth development (Koomen et al., 2012; Sass, 2011). Moreover, other classroom management observations systems have been developed with separate scales for different developmental levels, insinuating that classroom management practices vary across elementary, middle, and high school classrooms (Allen et al., 2013; Teachstone, 2011).

Practically, our results indicate that the ASSIST may be used to measure teachers' classroom management classroom in future applications, however, scores should appropriately account for differences in measurement properties across contexts by explicitly including MNI/DIF parameters in scoring. While our results do not suggest that a user can score the measure for one classroom context in the exact same manner as another classroom context, results do suggest that there are many items that are invariant across a given predictor (i.e., demonstrating partial invariance), allowing for meaningful

comparisons across contexts after the modeling and/or scoring accommodates for MNI/DIF (Vandenberg & Lance, 2000). These findings are especially helpful for districts interested in measuring teacher classroom practices similarly across elementary, middle, and high schools. This is also promising for researchers conducting large-scale descriptive or intervention studies who are searching for a measure of classroom management that can be used across school contexts.

Second, the results demonstrated that there were latent mean differences in latent scales across classroom contexts. Taken together, the differences tended to relate to developmental level, suggesting the similarity between the practices of elementary and middle school teachers; these results also suggested that practices of elementary school teachers differed from those of high school teachers. Specifically, teachers demonstrated greater proactive behavior support, monitoring, and responsiveness in elementary than high school classrooms. These findings suggest that teachers of younger students generally provided more direction and guidance to their students than teachers of older students. These results likely reflect the differing developmental abilities of these students; with elementary students not having the capacity for independence as do older students, and students experiencing a greater need for autonomy as they develop, it is reasonable that teachers' direction and monitoring peak in elementary school when students need the most behavioral support (Eccles et al., 1991; Mahatmya et al., 2012). Findings are also consistent with previous literature indicating that teachers of younger students utilize more positive praise than teachers of older students (Owens et al., 2018). However, latent mean differences also suggested that Disruptive Behavior was significantly greater in high school classrooms, as compared to elementary schools. This finding may suggest that high school teachers could benefit from the utilization of additional classroom management strategies.

Class Size

Only two significant findings emerged with respect to class size MNI/DIF, indicating that larger classes had a higher intercept for social conversations among students than smaller classes and a higher loading for the item "Students are irritable or sarcastic toward peers". As these were the only significant findings, we are hesitant to put too much emphasis on these results. Results largely indicated that

teachers' classroom management practices could be assessed in the same manner across classes of varying sizes. Although these results are consistent with studies reporting the measurement invariance of a self-report measure of students' substance use by school size (Babbin et al., 2011; Harrington et al., 2011; McGee et al., 2012), the current study is the first that we know of to assess measurement invariance of an observational measure of teachers' classroom practices by class size. This is an exceptionally important line of investigation considering the variability in class size (Finn & Achilles, 1999). It is likely that both researchers and practitioners would be interested in measuring classroom practices across classes of varying sizes. These results are again promising for districts, researchers, and practitioners interested in measuring teacher classroom practices across classroom contexts.

With respect to latent mean differences, results were contrary to previous studies in demonstrating that teachers were less responsive, monitored less, and provided less behavioral support in classes of larger sizes (Folmer-Annevelink et al., 2010; Hargreaves et al., 1998). Although the effect sizes in the present study were very small, they indicate that despite the potentially greater need for classroom management practices in larger classes, teachers may be struggling to *implement* effective strategies in these contexts.

Classroom Racial Composition

Similar to class size, there were few significant MNI/DIF findings related to classroom racial composition; only two items, both on the Responsiveness scale, had higher intercepts in classrooms in which a larger proportion of students were perceived by observers to be White. These two findings suggest some intercept invariance exists, with the implication being that, to use scores across settings, contextual differences across classrooms must be accounted for to create meaningful scores. However, given the partial invariance, it was possible to compare ASSIST scores across classrooms with varying racial compositions.

In terms of mean differences on the latent classroom management variables, classrooms with a higher proportion of students perceived to be White had significantly higher levels of teacher Influence and student Cooperation, and lower levels of student Disruptive Behavior. This is consistent with prior

research suggesting there are differences in teacher practices and student behavior based upon the racial composition of the classroom, such that in classrooms perceived to have more White students, teachers exhibit better classroom management and students exhibit more cooperative and less disruptive behavior (Authors, under review; Bottiani et al., 2019). However, given that we found partial invariance by racial composition, we can say with more confidence than in prior research that these findings are reflective of substantive differences rather than differences in measurement of the construct.

Limitations and Future Directions

There are many strengths of the present study, however, there are also several important limitations. First, despite pre-dating contemporary conceptualizations of classroom management (Collier-Meek et al., 2019), the global ASSIST scales capture foundational and preventative classroom management practices quite well. However, they provide a less comprehensive assessment of responsive classroom practices (i.e., only one scale captured responsive practices; Collier-Meek et al., 2019). Responsive practices are an important aspect of classroom management because they reinforce appropriate behavior and discourage problem behavior, a necessary complement to teachers' work in establishing clear expectations for students and preventing problems. Although the global scales were the focus of the present study, the ASSIST tally measures do capture responsive practices by assessing teachers' approvals and disapprovals of student behavior, and reactive behavior management. To fully assess the range of classroom management practices, researchers and practitioners may consider utilizing both the ASSIST global and tally measures.

Moreover, we delineated development according to classroom level: elementary, middle, and high school classrooms. Although it may be important to assess how measurement of these practices operates at a more nuanced grade level, we contrasted elementary, middle, and high school classrooms because we believe this grouping has a stronger practical application for those interested in administering the ASSIST. Additionally, although the sample size of teachers is quite large, the sample is entirely drawn from the Southern United States. This sample was not nationally representative, and thus, caution is warranted in generalizing findings to other populations. Similarly, all high school classrooms included in

the current research participated in one large-scale, statewide study (see Bradshaw et al., 2014). Additional replication research should be conducted to support partial MI and latent mean differences across elementary and high school classrooms.

We also took one particular approach to fitting a MNLFA, where we initially tested each item and covariate pair for significance (i.e., $p \leq .001$) before combining those significant predictors of DIF in a single model for each scale. However, some researchers have employed other approaches, such as item-by-item testing but with all covariate effects modeled together (Gottfredson et al., 2019). Alternatively, a traditional CFA MI testing framework could be employed, first testing for metric/weak invariance and subsequently scalar/strong invariance (e.g., Vandenberg & Lance, 2000). It is possible that alternative procedures may have produced a different pattern of results.

We also explored variation based on class size and racial composition on DIF and latent means, but we did not explore the extent to which the findings varied as a function of other contextual factors (e.g., teachers' gender, subject, class achievement levels). Exploration of these and other possible sources of systematic variance, both with regard to MI and mean level differences in ASSIST scale scores, should be explored in future research. In particular, it is important to reiterate that in the current study, teachers were rated by one observer on a single occasion. As such, we were not able to account for teacher practices across occasions and our analytic approach does not account for variance attributable to raters. Future research on the ASSIST should explore these potentially important factors (Casabianca et al., 2015). For example, Briggs and Alzen (2019) determined that it would be necessary to observe teachers eight times over the course of two years to reliably capture change in teacher practices over time. Additional research is needed to discern the optimal number of observations needed for specific types of use (e.g., point in time versus change over time). However, in past research using the ASSIST, ratings were averaged over multiple observation time points, as scores across multiple ASSIST observations demonstrated very high intraclass correlations at the teacher level (i.e., item-level intraclass correlations ranged from .72 to .81), indicating relatively little variability across the within-teacher repeated observations (Cicchetti, 1994).

Moreover, our findings suggested that there were latent mean differences in classroom constructs across developmental periods. Additional research is necessary to determine whether higher levels of these constructs are universally promotive of student outcomes across development, or whether specific profiles of teacher and student behaviors are optimal for promoting students' social, emotional, behavioral, and academic outcomes in certain developmental periods. Finally, as noted in the discussion, standardized expert observations can be burdensome and expensive for schools to implement outside of large-scale research trials (Kane & Staiger, 2012; Rowan & Correnti, 2009). Researchers have provided suggestions for less costly alternatives such as teacher-reported logs of behavior (Rowan & Correnti, 2009) and student-report (Kane & Staiger, 2012). Student-reported measures of teacher and classroom practices have been increasingly utilized (Kane & Staiger, 2012; Wallace et al., 2016). However, it is posited that students have a different frame of reference for classroom practices, as compared to expert observers and teachers themselves, and work aimed at understanding the theoretical frameworks that underly student reports of teacher and classroom practices is still ongoing (Wallace et al., 2016). Moreover, given the utility of expert observation for performance feedback and coaching (e.g., Reinke et al., 2011, 2015), we believe the benefits offset the burden. Regardless, we see great value in using multiple methods to provide a more comprehensive view of teachers' classroom management practices.

Conclusions and Implications

The current study provides encouraging findings for researchers and practitioners by indicating that the ASSIST observational measure of teachers' classroom management practices may be used across various classroom contexts, so long as MNI/DIF is accounted for, yielding meaningful comparisons across these contexts. Most of the MNI/DIF was attributable to differences in measurement across developmental levels, with less frequent MNI/DIF noted across class size and racial composition. Differences in latent means across elementary, middle, and high school classrooms may highlight the need for additional professional development and coaching supports for teachers in high school classrooms, where teachers generally used fewer directive and praise-based strategies, compared to elementary teachers. Such findings suggest that secondary teachers may benefit from additional coaching

related to classroom management strategies, whereas the tendency is often to focus more on elementary school teachers in such supports (Owens et al., 2018). With regard to future studies, the ASSIST appears to be robust with regard to partial MI, and thus holds great promise as a tool for consistently measuring teachers' use of effective classroom management practices and corresponding student behavior across classroom context and developmental level. As such, the ASSIST could serve as a single tool to be used with confidence across multiple classroom settings and across school levels. In turn, stakeholders using the ASSIST may draw meaningful comparisons of teachers' use of classroom management strategies across elementary, middle, and high school settings.

References

Aguiar, A. L., & Aguiar, C. (2020). Classroom composition and quality in early childhood education: A systematic review. *Children and Youth Services Review*, *115*, 1–26. https://doi.org/10.1016/j.childyouth.2020.105086

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review*, *42*(1), 76–98. https://doi.org/10.1080/02796015.2013.12087492

Babbin, S. F., Harrington, M., Burditt, C., Redding, C., Paiva, A., Meier, K., Oatley, K., McGee, H., & Velicer, W. F. (2011). Prevention of alcohol use in middle school students: Psychometric assessment of the decisional balance inventory. *Addictive Behaviors*, *36*(5), 543–546. https://doi.org/10.1016/j.addbeh.2011.01.010

Barrett, S. B., Bradshaw, C. P., & Lewis-Palmer, T. (2008). Maryland statewide PBIS initiative: Systems, evaluation, and next steps. *Journal of Positive Behavior Interventions*, *10*(2), 105–114. https://doi.org/10.1177/1098300707312541

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101-125. https://doi.org/10.1037/a0015583

Betts, J. E., Appleton, J. J., Reschly, A. L., Christenson, S. L., & Huebner, E. S. (2010). A study of the factorial invariance of the Student Engagement Instrument (SEI): Results from middle and high school students. *School Psychology Quarterly*, *25*(2), 84–93. https://doi.org/10.1037/a0020259

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and

primary vs. secondary schools. *Learning and Instruction*, *21*(6), 715–730. https://doi.org/10.1016/j.learninstruc.2011.04.001

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), S176–S181.

Bottiani, J. H., Duran, C. A. K., Pas, E. T., & Bradshaw, C. P. (2019). Teacher stress and burnout in urban middle schools: Associations with job demands, resources, and effective classroom practices. *Journal of School Psychology*, *77*, 36–51. https://doi.org/10.1016/j.jsp.2019.10.002

Bourke, S. (1986). How smaller is better: Some relationships between class size, teaching practices, and student achievement. *American Educational Research Journal*, *23*(4), 558–571. https://doi.org/10.3102/00028312023004558

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide Positive Behavioral Interventions and Supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148. https://doi.org/10.1177/1098300709334798

Bradshaw, C. P., & Pas, E. T. (2011). A statewide scale up of positive behavioral interventions and supports: A description of the development of systems of support and analysis of adoption and implementation. *School Psychology Review*, *40*(4), 530–548. https://doi.org/10.1080/02796015.2011.12087528

Bradshaw, C. P., Pas, E. T., Bloom, J., Barrett, S., Hershfeldt, P., Alexander, A., McKenna, M., Chafin, A. E., & Leaf, P. J. (2012). A state-wide partnership to promote safe and supportive schools: The PBIS Maryland Initiative. *Administration and Policy in Mental Health and Mental Health Services Research*, *39*(4), 225–237. https://doi.org/10.1007/s10488-011-0384-6

Bradshaw, C. P., Pas, E. T., Bottiani, J. H., Debnam, K. J., Reinke, W. M., Herman, K. C., & Rosenberg, M. S. (2018). Promoting cultural responsivity and student engagement through Double Check coaching of classroom teachers: An efficacy study. *School Psychology Review*, *47*(2), 118–134. https://doi.org/10.17105/SPR-2017-0119.V47-2

Bradshaw, C. P., Pas, E. T., Debnam, K. J., & Johnson, S. L. (2021). A randomized controlled trial of MTSS-B in high schools: Improving classroom management to prevent EBDs. *Remedial and Special Education*, *42*(1), 44–59. https://doi.org/10.1177/0741932520966727

Bradshaw, C. P., Waasdorp, T. E., Debnam, K. J., & Johnson, S. L. (2014). Measuring school climate in high schools: A focus on safety, engagement, and the environment. *Journal of School Health*, *84*(9), 593–604. https://doi.org/10.1111/josh.12186

Briggs, D. C., & Alzen, J. L. (2019). Making inferences about teacher observation scores over time. *Educational and Psychological Measurement*, *79*(4), 636–664. https://doi.org/10.1177/0013164419826237

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.

Bronfenbrenner, U., & Morris, P. A. (2007). The bioecological model of human development. In W. Damon & R.M. Lerner (Eds.), *Handbook of Child Psychology (*pp. 793 – 828). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0114

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456. https://doi.org/10.1037/0033-2909.105.3.456

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311–337. https://doi.org/10.1177/0013164414539163

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

Clunies-Ross, P., Little, E., & Kienhuis, M. (2008). Self-reported and actual use of proactive and reactive classroom management strategies and their relationship with teacher stress and student behaviour. *Educational Psychology*, *28*(6), 693–710. https://doi.org/10.1080/01443410802206700

Collier-Meek, M. A., Johnson, A. H., Sanetti, L. H., & Minami, T. (2019). Identifying critical components of classroom management implementation. *School Psychology Review*, *48*(4), 348–361. https://doi.org/10.17105/SPR-2018-0026.V48-4

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Curran, P., & Hussong, A. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Debnam, K. J., Pas, E. T., Bottiani, J., Cash, A. H., & Bradshaw, C. P. (2015). An examination of the association between observed and self-reported culturally proficient teaching practices. *Psychology in the Schools*, *52*(6), 533–548. https://doi.org/10.1002/pits.21845

Downer, J., Sabol, T. J., & Hamre, B. (2010). Teacher–child interactions in the classroom: Toward a theory of within-and cross-domain links to children's developmental outcomes. *Early Education and Development*, *21*(5), 699-723.

Downer, J. T., López, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher–child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, *27*(1), 21–32. https://doi.org/10.1016/j.ecresq.2011.07.005

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, *77*(4), 267–282. https://doi.org/10.1177/003804070407700401

Eccles, J. S., Buchanan, C. M., Flanagan, C., Fuligni, A., Midgley, C., & Yee, D. (1991). Control versus autonomy during early adolescence. *Journal of Social Issues*, *47*(4), 53–68. https://doi.org/10.1111/j.1540-4560.1991.tb01834.x

Eccles, J. S., & Roeser, R. W. (2011). School and community influences on human development. In M. H. Bornstein & M. E. Lamb (Eds.), *Developmental science: An advanced textbook, 6th ed* (pp. 571–643). Psychology Press.

Evertson, C. M., & Randolph, C. H. (1989). Teaching practices and class size: A new look at an old issue. *Peabody Journal of Education*, *67*(1), 85–105. https://doi.org/10.1080/01619569209538671

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*(3), 557–577. https://doi.org/10.3102/00028312027003557

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, *21*(2), 97–109. https://doi.org/10.3102/01623737021002097

Folmer-Annevelink, E., Doolaard, S., Mascareño, M., & Bosker, R. J. (2010). Class size effects on the number and types of student-teacher interactions in primary classrooms. *The Journal of Classroom Interaction*, *45*(2), 30–38.

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

Gaertner, H., & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(2), 159–182. https://doi.org/10.1007/s11092-018-9277-5

Gaias, L. M., Lindstrom Johnson, S., Bottiani, J. H., Debnam, K. J., & Bradshaw, C. P. (2019). Examining teachers' classroom management profiles: Incorporating a focus on culturally responsive practice. *Journal of School Psychology*, *76*, 124–139. https://doi.org/10.1016/j.jsp.2019.07.017

Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., & Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: Automated moderated nonlinear factor analysis (aMNLFA). *Addictive Behaviors*, *94*, 65-73. https://doi.org/10.1016/j.addbeh.2018.10.031

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–83). Paul H Brookes Publishing.

Hargreaves, L., Galton, M., & Pell, A. (1998). The effects of changes in class size on teacher–pupil interaction. *International Journal of Educational Research*, *29*(8), 779–795. https://doi.org/10.1016/S0883-0355(98)00063-9

Harrington, M., Babbin, S., Redding, C., Burditt, C., Paiva, A., Meier, K., Oatley, K., McGee, H., & Velicer, W. (2011). Psychometric assessment of the Temptations to Try Alcohol Scale. *Addictive Behaviors*, *36*(4), 431–433. https://doi.org/10.1016/j.addbeh.2010.12.019

Hickey, G., McGilloway, S., Hyland, L., Leckey, Y., Kelly, P., Bywater, T., Comiskey, C., Lodge, A., Donnelly, M., & O'Neill, D. (2017). Exploring the effects of a universal classroom management training programme on teacher and child behaviour: A group randomised controlled trial and cost analysis. *Journal of Early Childhood Research*, *15*(2), 174–194. https://doi.org/10.1177/1476718X15579747

Hirsch, S. E., Randall, K., Bradshaw, C., & Lloyd, J. W. (2021). Professional learning and development in classroom management for novice teachers: A systematic review. *Education and Treatment of Children*, *44*(4), 291–307. https://doi.org/10.1007/s43494-021-00042-6

Horner, R. H., Kincaid, D., Sugai, G., Lewis, T., Eber, L., Barrett, S., Dickey, C. R., Richter, M., Sullivan, E., Boezio, C., Algozzine, B., Reynolds, H., & Johnson, N. (2014). Scaling up School-Wide Positive Behavioral Interventions and Supports: Experiences of seven states with documented success. *Journal of Positive Behavior Interventions*, *16*(4), 197–208. https://doi.org/10.1177/1098300713503685

Kalender, İ. (2015). Measurement invariance of student evaluation of teaching across groups defined by course-related variables. *International Online Journal of Educational Sciences, 7*(4), 69–79. https://doi.org/10.15345/iojes.2015.04.006

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. In *Bill & Melinda Gates Foundation*. Bill & Melinda Gates Foundation. https://eric.ed.gov/?id=ED540960

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.

Koomen, H. M. Y., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2012). Validating the Student-Teacher Relationship Scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology*, *50*(2), 215–234. https://doi.org/10.1016/j.jsp.2011.09.001

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643–680. https://doi.org/10.3102/0034654315626799

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal*, *104*(5), 409–426. https://doi.org/10.1086/499760

Larson, K. E., & Bradshaw, C. P. (2017). Cultural competence and social desirability among practitioners: A systematic review of the literature. *Children and Youth Services Review*, *76*, 100–111. https://doi.org/10.1016/j.childyouth.2017.02.034

Larson, K. E., Pas, E. T., Bottiani, J. H., Kush, J. M., & Bradshaw, C. P. (2021). A multidimensional and multilevel examination of student engagement and secondary school teachers' use of classroom

management practices. *Journal of Positive Behavior Interventions*, 1098300720929352. https://doi.org/10.1177/1098300720929352

Larson, K. E., Pas, E. T., Bradshaw, C. P., Rosenberg, M. S., & Day-Vines, N. L. (2018). Examining how proactive management and culturally responsive teaching relate to student behavior: Implications for measurement and practice. *School Psychology Review*, *47*(2), 153–166. https://doi.org/10.17105/SPR-2017-0070.V47-2

Mahatmya, D., Lohman, B. J., Matjasko, J. L., & Farb, A. F. (2012). Engagement across developmental periods. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 45–63). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_3

Martin, P. A., Daley, D., Hutchings, J., Jones, K., Eames, C., & Whitaker, C. J. (2010). The Teacher-Pupil Observation Tool (T-POT): Development and testing of a new classroom observation measure. *School Psychology International*, *31*(3), 229–249. https://doi.org/10.1177/0143034310362040

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749. https://doi.org/10.1111/j.1467-8624.2008.01154.x

McGee, H. A., Babbin, S. F., Redding, C., Paiva, A., Oatley, K., Meier, K. S., Harrington, M., & Velicer, W. F. (2012). Prevention of smoking in middle school students: Psychometric assessment of the Temptations to Try Smoking Scale. *Addictive Behaviors*, *37*(4), 521–523. https://doi.org/10.1016/j.addbeh.2011.11.013

Mosteller, F. (1995). The tennessee study of class size in the early school grades. *The Future of Children*, *5*(2), 113–127. https://doi.org/10.2307/1602360

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. JSTOR. https://doi.org/10.2307/271070

Muthén, L. K., & Muthen, B. (1998). *Mplus user's guide: Statistical analysis with latent variables, user's guide*. Muthén & Muthén.

Oliver, R. M., Wehby, J. H., & Reschly, D. J. (2011). Teacher classroom management practices: Effects on disruptive or aggressive student behavior. *Campbell Systematic Reviews, 7*(1), 1–55. https://doi.org/10.4073/csr.2011.4

Oude Groote Beverborg, A., Feldhoff, T., Maag Merki, K., & Radisch, F. (Eds.). (2021). *Concept and Design Developments in School Improvement Research: Longitudinal, Multilevel and Mixed Methods and Their Relevance for Educational Accountability*. Springer Nature. https://doi.org/10.1007/978-3-030-69345-9

Owens, J. S., Holdaway, A. S., Smith, J., Evans, S. W., Himawan, L. K., Coles, E. K., Girio-Herrera, E., Mixon, C. S., Egan, T. E., & Dawson, A. E. (2018). Rates of common classroom behavior management strategies and their associations with challenging student behavior in elementary school. *Journal of Emotional and Behavioral Disorders, 26*(3), 156–169. https://doi.org/10.1177/1063426617712501

Pas, E. T., Bradshaw, C. P., & Mitchell, M. M. (2011). Examining the validity of office discipline referrals as an indicator of student behavior problems. *Psychology in the Schools, 48*(6), 541–555. https://doi.org/10.1002/pits.20577

Pas, E. T., Cash, A. H., O'Brennan, L., Debnam, K. J., & Bradshaw, C. P. (2015). Profiles of classroom behavior in high schools: Associations with teacher behavior management strategies and classroom composition. *Journal of School Psychology, 53*(2), 137–148. https://doi.org/10.1016/j.jsp.2014.12.005

Pas, E.T., Ryoo, J.H., Musci, R., & Bradshaw, C. P. (2019). The effects of a state-wide scale-up of school-wide Positive Behavior Intervention and Supports on behavioral and academic outcomes: A quasi-experimental examination. Journal of School Psychology, 73(1), 41-55. https://doi.org/10.1016/j.jsp.2019.03.001.

Pas, E. T., Waasdorp, T. E., & Bradshaw, C. P. (2019). Coaching teachers to detect, prevent, and respond to bullying using mixed reality simulation: An efficacy study in middle schools. *International Journal of Bullying Prevention*, *1*(1), 58–69. https://doi.org/10.1007/s42380-018-0003-0

Pianta, R. C. (2001). *Student-Teacher Relationship Scale: Professional manual*. Psychological Assessment Resources.

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

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431–451. https://doi.org/10.1016/j.ecresq.2008.02.001

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review: DR*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Reinke, W. M., Herman, K. C., & Sprick, R. (2011). *Motivational interviewing for effective classroom management: The Classroom Check-up*. Guilford Press.

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

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher*, *38*(2), 120–131. https://doi.org/10.3102/0013189X09332375

Rusby, J. C., Crowley, R., Sprague, J., & Biglan, A. (2011). Observations of the middle school environment: The context for student behavior beyond the classroom. *Psychology in the Schools*, *48*(4), 400–415. https://doi.org/10.1002/pits.20562

Rusby, J. C., Taylor, T., & Milchak, C. (2001). *Assessing school settings: Interactions of students and teachers (ASSIST) observation system* [Unpublished manual]. Oregon Research Institute.

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347–363. https://doi.org/10.1177/0734282911406661

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317

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

Solorzano, D. G. (1997). Images and words that wound: Critical race theory, racial stereotyping, and teacher education. *Teacher Education Quarterly*, *24*(3), 5–19.

Stichter, J. P., Lewis, T. J., Johnson, N., & Trussell, R. (2004). Toward a structural assessment: Analyzing the merits of an assessment tool for a student with E/BD. *Assessment for Effective Intervention*, *30*(1), 25–40. https://doi.org/10.1177/073724770403000103

Stichter, J. P., Lewis, T. J., Whittaker, T. A., Richter, M., Johnson, N. W., & Trussell, R. P. (2009). Assessing teacher use of opportunities to respond and effective classroom management strategies: Comparisons among high- and low-risk elementary schools. *Journal of Positive Behavior Interventions*, *11*(2), 68–81. https://doi.org/10.1177/1098300708326597

Teachstone. (2011). *The CLASS® System: Birth through Secondary*. Teachstone. http://info.teachstone.com/hubfs/CLASS-System-Paper-FINAL.pdf?t=1495735615702

Tolan, P., Elreda, L. M., Bradshaw, C. P., Downer, J. T., & Ialongo, N. (2020). Randomized trial testing the integration of the Good Behavior Game and MyTeachingPartner™: The moderating role of distress among new teachers on student outcomes. *Journal of School Psychology*, *78*, 75–95. https://doi.org/10.1016/j.jsp.2019.12.002

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Waasdorp, T. E., Lindstrom Johnson, S., Shukla, K. D., & Bradshaw, C. P. (2020). Measuring school climate: Invariance across middle and high school students. *Children & Schools*, *42*(1), 53–62. https://doi.org/10.1093/cs/cdz026

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, *53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Watts-Taffe, S., Laster, B. P. (Barbara), Broach, L., Marinak, B., Connor, C. M., & Walker-Dalhouse, D. (2012). Differentiated instruction: Making informed teacher decisions. *The Reading Teacher*, *66*(4), 303–314. https://doi.org/10.1002/TRTR.01126

What Works Clearinghouse. (2020). *What Works Clearinghouse Standards Handbook, version 4.1*. US Department of Education, Institute of Education Sciences. National Center for Education Evaluation and Regional Assistance. https://ies. ed. gov/ncee/wwc/Docs/ref-erenceresources/WWC-Standards-Handbook-v4-1-508. pdf.

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2020). *dplyr: A Grammar of Data Manipulation* (1.0.2) [Computer software]. https://CRAN.R-project.org/package=dplyr

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction, 66,* 101303. https://doi.org/10.1016/j.learninstruc.2020.101303

Yu, C.-Y. & Muthén, B. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA. April.

Table 1. *Teacher and Classroom Descriptive Statistics by Developmental Level*

	Elementary (*n* = 616)		**Middle (*n* = 1,256)**		**High (*n* =1,391)**	
Teacher	***%***		***%***		***%***	
Teacher Gender [a]						
Male	6.82		7.56		39.76	
Female	82.14		21.26		60.24	
Students	***M (SD)***	***Range***	***M (SD)***	***Range***	***M (SD)***	***Range***
Number of Students	20.31 (4.85)	2 – 50	23.59 (5.93)	3 - 43	18.59 (6.46)	1 - 45
Percentage Male	49.66 (10.61)	0 - 88.89	50.65 (14.03)	0 - 100.00	49.70 (16.56)	0 - 100.00
Percentage White	34.73 (20.73)	0 - 92.00	33.48 (26.34)	0 - 100.00	52.33 (30.23)	0 - 100.00

Note. [a] Percentages not summing to 100% indicate the percentage of missing data.

Table 2. *Teacher ASSIST Item-level Descriptive Statistics*

		Elementary				Middle				High			
	Item	***M***	***SD***	***Min***	***Max***	***M***	***SD***	***Min***	***Max***	***M***	***SD***	***Min***	***Max***
Influence	Teacher and students appear comfortable with one another	3.40	0.73	0.00	4.00	3.46	0.76	0.00	4.00	3.37	0.84	0.00	4.00
	Teacher has good control of or influence on students	3.24	0.82	0.00	4.00	3.11	0.95	0.00	4.00	3.08	0.93	0.00	4.00
	Teacher has little/no control of or influence on students	0.70	1.05	0.00	4.00	0.73	0.98	0.00	4.00	0.92	0.96	0.00	4.00
	There are instances of teacher annoyance, irritability, or sarcasm directed at students	0.54	0.96	0.00	4.00	0.40	0.74	0.00	4.00	0.39	0.69	0.00	4.00
	There is evidence of classroom routines (students know what they're supposed to be doing)	3.37	0.74	0.00	4.00	3.33	0.81	0.00	4.00	3.32	0.82	0.00	4.00
Behavior Support	Teacher gives clear instructions and directives to students.	3.32	0.78	0.00	4.00	3.44	0.73	0.00	4.00	3.17	0.93	0.00	4.00
	Teacher is consistent, even-handed, and firm when necessary	3.29	0.82	0.00	4.00	3.28	0.89	0.00	4.00	2.88	1.15	0.00	4.00
	Teacher praised students for specific behaviors or using social skills	2.14	1.25	0.00	4.00	1.88	1.33	0.00	4.00	1.55	1.25	0.00	4.00
	The teacher clearly explains learning objectives prior to and/or during the lesson through summary or re-orientation statements	2.84	1.26	0.00	4.00	2.98	1.19	0.00	4.00	2.76	1.12	0.00	4.00
Monitoring	Teachers is able to focus on one or two students while still scanning all other areas.	3.08	0.96	0.00	4.00	2.95	1.05	0.00	4.00	3.05	0.97	0.00	4.00
	Teacher monitors all students and all areas	3.29	0.83	0.00	4.00	3.25	0.84	0.00	4.00	3.20	0.89	0.00	4.00
	Teacher positions him/herself so they can see most of the room area	3.45	0.74	0.00	4.00	3.44	0.76	0.00	4.00	3.44	0.81	0.00	4.00
	Teacher scans the room and is aware of what is occurring	3.35	0.80	0.00	4.00	3.29	0.83	0.00	4.00	3.22	0.86	0.00	4.00
Meaningful Participation	Teacher encourages students to share their ideas and opinions	2.26	1.39	0.00	4.00	2.39	1.38	0.00	4.00	2.76	1.12	0.00	4.00
	Teacher gets students involved in lesson by asking questions or making comments	2.94	1.06	0.00	4.00	2.67	1.25	0.00	4.00	2.81	1.14	0.00	4.00
	Students are provided opportunities to contribute to discussion	2.56	1.24	0.00	4.00	2.35	1.37	0.00	4.00	2.58	1.20	0.00	4.00
	Students have opportunities to make choices	1.70	1.38	0.00	4.00	1.57	1.45	0.00	4.00	2.00	1.40	0.00	4.00
	Students have opportunities to take leadership roles in the classroom	1.24	1.37	0.00	4.00	0.99	1.30	0.00	4.00	1.38	1.37	0.00	4.00
	Students are prosocial toward one another	2.32	1.23	0.00	4.00	2.10	1.26	0.00	4.00	2.26	1.20	0.00	4.00
	Students praise and compliment one another	0.89	1.32	0.00	4.00	0.85	1.21	0.00	4.00	1.46	1.29	0.00	4.00
	Students respond to teacher's questions and/or volunteer when asked	3.24	0.98	0.00	4.00	3.21	0.99	0.00	4.00	3.05	1.01	0.00	4.00
	Students share their ideas/opinions	1.53	1.42	0.00	4.00	1.54	1.36	0.00	4.00	2.62	1.05	0.00	4.00
Responsiveness	Teacher anticipates when students may have problems academically	2.82	1.07	0.00	4.00	2.74	1.20	0.00	4.00	2.81	1.11	0.00	4.00
	Teacher anticipates when students may have problems behaviorally	2.88	1.00	0.00	4.00	2.58	1.23	0.00	4.00	2.40	1.23	0.00	4.00
	Teacher is responsive to students' behavioral and/or emotional needs	3.17	0.83	0.00	4.00	2.90	1.03	0.00	4.00	2.80	1.07	0.00	4.00
	Teacher maintains proximity to students who display a need for assistance or support[a]	3.30	0.86	1.00	4.00	2.99	1.05	0.00	4.00	2.71	1.07	0.00	4.00
	Teacher notices when students have difficulty understanding a concept	2.88	1.07	0.00	4.00	2.99	1.00	0.00	4.00	3.03	0.95	0.00	4.00
	Teacher uses verbal reminders or nonverbal cues regarding expected behaviors	2.75	1.08	0.00	4.00	2.55	1.16	0.00	4.00	2.58	1.16	0.00	4.00

Note. [a] The full range of response options (i.e., 0 – 4) was not endorsed for this item across elementary, middle, and high school classrooms.

Table 3. *Student ASSIST Item-level Descriptive Statistics*

		Elementary				Middle				High			
	Item	***M***	***SD***	***Min***	***Max***	***M***	***SD***	***Min***	***Max***	***M***	***SD***	***Min***	***Max***
Cooperation	Students are focused and engaged	3.08	0.82	0.00	4.00	2.92	0.92	0.00	4.00	3.01	0.92	0.00	4.00
	Students are interested, enthusiastic, and involved	3.05	0.84	0.00	4.00	2.86	0.95	0.00	4.00	2.86	0.97	0.00	4.00
	Students comply [a]	3.22	0.77	1.00	4.00	3.19	0.82	0.00	4.00	3.18	0.81	0.00	4.00
	Students consistently follow rules appropriate to settings	3.19	0.79	0.00	4.00	3.10	0.85	0.00	4.00	3.16	0.83	0.00	4.00
	Students cooperate [a]	3.27	0.73	1.00	4.00	3.25	0.81	0.00	4.00	3.28	0.78	0.00	4.00
	Students treat their peers with respect	2.91	0.97	0.00	4.00	2.97	0.95	0.00	4.00	3.03	0.93	0.00	4.00
	Students handle transitions well	2.82	1.26	0.00	4.00	3.00	0.93	0.00	4.00	2.95	0.93	0.00	4.00
Disruptive Behavior	Social conversations occur between students and peers	1.87	1.34	0.00	4.00	2.50	1.34	0.00	4.00	1.89	1.12	0.00	4.00
	Students are irritable or sarcastic toward peers	0.48	0.84	0.00	4.00	0.30	0.75	0.00	4.00	0.17	0.53	0.00	4.00
	Students are irritable or sarcastic toward the teacher	0.21	0.56	0.00	4.00	0.14	0.48	0.00	4.00	0.11	0.46	0.00	4.00
	Students argue with peers	0.41	0.75	0.00	4.00	0.13	0.49	0.00	4.00	0.04	0.27	0.00	4.00
	Students argue with the teacher [a]	0.15	0.45	0.00	3.00	0.10	0.39	0.00	3.00	0.05	0.29	0.00	4.00
	Students engage in verbal aggression toward teachers [a]	0.08	0.36	0.00	4.00	0.01	0.15	0.00	3.00	0.00	0.06	0.00	1.00
	Students physically harass and/or bully others [a]	0.19	0.54	0.00	4.00	0.06	0.30	0.00	3.00	0.02	0.15	0.00	3.00

Note. [a] The full range of response options (i.e., 0 – 4) was not endorsed for this item across elementary, middle, and high school classrooms.

Table 4. *Final Multivariate MNLFA Results*

		Loading DIF							Intercept DIF			
Factor/Item	Loading	Size	MS vs. ES	HS vs. ES	T1	T2	T3	T4	Size	% White	MS vs. ES	HS vs. ES
Influence and Cooperation (n = 3,247)												
Teacher and students appear comfortable with one another	1.81				-7.95	-5.83	-3.69	-1.23				
Teacher has good control of or influence on students	3.45				-11.04	-7.99	-4.51	-0.71				
Teacher has little/no control of or influence on students	-2.14				1.27	3.66	5.63	7.48				**0.79***
There are instances of teacher annoyance, irritability, or sarcasm directed at students	-1.14				1.64	3.09	4.82	6.44				
There is evidence of classroom routines (students know what they're supposed to be doing)	2.46				-9.15	-7.33	-4.32	-1.17				
Students are focused and engaged	3.89				-12.68	-9.00	-4.23	0.09				
Students are interested, enthusiastic, and involved	2.52				-8.56	-6.03	-2.48	0.33				
Students comply	4.60				-11.96	-6.62	-0.95					
Students consistently follow rules appropriate to settings	4.85				-17.41	-12.13	-6.41	-0.78				
Students cooperate	5.70				-15.51	-8.64	-1.91					
Students treat their peers with respect	2.29				-7.38	-5.79	-2.76	0.10				
Students handle transitions well	2.00				-5.84	-4.72	-2.52	-0.15				
Behavior Support (n = 3,247)												
Teacher gives clear instructions and directives to students	3.60				-11.83	-9.61	-6.25	-2.08				
Teacher is consistent, even-handed, and firm when necessary	1.97				-6.72	-4.89	-3.04	-0.86				
Teacher praised students for specific behaviors or using social skills	0.96				-2.13	-1.04	0.49	1.68				
The teacher clearly explains learning objectives prior to and/or during the lesson through summary or re-orientation statements	2.41			**-0.50***	-5.62	-4.46	-2.59	-0.45				
Monitoring (*n* = 3,247)												
Teachers is able to focus on one or two students while still scanning all other areas.	1.79		**0.61***		-6.76	-5.09	-2.65	-0.27				
Teacher monitors all students and all areas	6.38				-20.26	-15.26	-9.01	-2.73				
Teacher positions him/herself so they can see most of the room area	3.17				-11.84	-9.01	-5.67	-2.46				
Teacher scans the room and is aware of what is occurring	6.32				-21.01	-15.73	-9.14	-3.02				

Meaningful Engagement (n = 3,245)											
Teacher encourages students to share their ideas and opinions	4.80			-0.61	-8.84	-7.00	-3.87	-0.64			
Teacher gets students involved in lesson by asking questions or making comments	2.85				-7.40	-5.58	-3.16	-1.05			**-0.50***
Students are provided opportunities to contribute to discussion	3.38				-6.83	-5.27	-2.74	-0.19			
Students have opportunities to make choices	1.58				-2.43	-1.59	-0.19	1.19			
Students have opportunities to take leadership roles in the classroom	1.60				-1.40	-0.57	0.90	2.00			
Students are prosocial toward one another	0.87				-2.80	-1.82	-0.14	1.13			
Students praise and compliment one another	1.30				-0.69	0.25	1.74	2.71			**0.92***
Students share their ideas/opinions	2.38				-3.35	-2.23	0.06	2.26			**1.91***
Responsiveness (*n* = 3,247)											
Teacher anticipates when students may have problems academically	2.89				-7.65	-5.97	-3.37	-0.64		**0.11***	
Teacher anticipates when students may have problems behaviorally	2.74				-6.78	-5.09	-2.45	-0.04			
Teacher is responsive to students' behavioral and/or emotional needs	2.38				-7.65	-5.63	-3.04	-0.56			
Teacher maintains proximity to students who display a need for assistance or support	1.24		**0.98***		-4.91	-2.84	-0.95				**-1.29***
Teacher notices when students have difficulty understanding a concept	2.77				-8.44	-6.35	-3.71	-0.85		**0.14***	
Teacher uses verbal reminders or nonverbal cues regarding expected behaviors	1.24				-4.31	-3.10	-1.23	0.36			
Disruptive Behavior (*n* = 3,263)											
Social conversations occur between students and peers	0.42		**1.01***		-1.70	-0.33	1.28	2.35	**0.05***		
Students are irritable or sarcastic toward peers	3.15	**0.03***			6.47	8.07	10.90	12.82			
Students are irritable or sarcastic toward the teacher	1.99				5.38	6.78	8.75	10.03			
Students argue with peers	4.38				10.60	12.68	16.29	19.39			
Students argue with the teacher	1.98				6.04	7.53	9.65				
Students engage in verbal aggression toward teachers	1.81				7.43	8.94					
Students physically harass and/or bully others	2.61				7.90	9.76	11.75				

Notes. Classroom racial composition was omitted as a predictor of loading DIF, as it did not yield statistically significant impacts on item loadings; DIF parameter estimates that are statistically significant at the $p < .001$ level are denoted in bold and with an asterisk.

Table 5. *Predictors of ASSIST Scale Factor Levels*

		Predictors			
ASSIST Scale		*Size*	*% White*	*MS vs. ES*	*HS vs. ES*
Influence and Cooperation	*b*	-0.02	0.06	-0.07	-0.55
	S.E.	0.00	0.02	0.19	0.18
	p	< .001	< .001	0.718	0.002
Behavior Support	*b*	-0.02	0.02	0.00	-0.90
	S.E.	0.01	0.02	0.20	0.22
	p	0.001	0.209	0.996	< .001
Monitoring	*b*	-0.02	0.04	-0.22	-0.64
	S.E.	0.01	0.02	0.14	0.15
	p	< .001	0.009	0.127	< .001
Meaningful Engagement	*b*	-0.01	0.02	0.00	-0.59
	S.E.	0.01	0.02	0.30	0.30
	p	0.033	0.254	0.99	0.045
Responsiveness	*b*	-0.03	0.02	0.07	-0.89
	S.E.	0.01	0.02	0.23	0.23
	p	< .001	0.314	0.780	< .001
Disruptive Behavior	*b*	0.00	-0.10	0.45	0.64
	S.E.	0.01	0.02	0.16	0.19
	p	0.472	< .001	0.005	0.001