12-2022

A psychometric analysis of the comprehensive school survey (CSS) middle school student version.

Stephen Michael Leach
University of Louisville

A PSYCHOMETRIC ANALYSIS OF THE COMPREHENSIVE SCHOOL SURVEY (CSS) MIDDLE SCHOOL STUDENT VERSION

By

Stephen Michael Leach
B.A., University of Louisville, 2010

A Dissertation Submitted to the Faculty of the
College of Education and Human Development of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
In Counseling and Personnel Services

Department of
Educational Psychology, Measurement, & Evaluation
University of Louisville

December 2022

A PSYCHOMETRIC ANALYSIS OF THE COMPREHENSIVE SCHOOL SURVEY (CSS) MIDDLE SCHOOL STUDENT VERSION

By

Stephen Michael Leach
B.A., University of Louisville, 2010

A Dissertation Approved on

November 11, 2022

by the following Dissertation Committee:

Dr. Jeffrey C. Valentine, Co-Chair

Dr. Jason C. Immekus, Co-Chair

Dr. Prathiba Natesan Batley

Dr. Dena Dossett

DEDICATION

To my daughter, Deborah Miele Leach, who, at the time of my studies,
cared not a whit for psychometric investigations or instrument design,
but loved art and soccer and hiking the many trails in Iroquois Park
and reading about young wizards and Narnian kings and queens.
May you be ever curious and creative, amazed by creation,
your heart filled with love and joy and wonder.

The LORD bless you and keep you;
the LORD make his face to shine upon you and be gracious to you;
the LORD lift up his countenance upon you and give you peace.
Numbers 6:24-26

ACKNOWLEDGEMENTS

This culminating effort of my prolonged collegiate odyssey was by no means a solo effort. I received support, encouragement, and cajoling along the way from more people than I can mention here. Thanks goes first to the Lord, my refuge and strength in time of need, and then to my mom, dad, and stepmom, who stuck by me in my wayward days, when life was unmanageable and going to college was not even an afterthought.

I owe an immense debt of gratitude to my dissertation committee, especially Drs. Jeff Valentine and Jason Immekus, whose wisdom, patience, encouragement, availability, and commitment to excellence were invaluable throughout my doctoral studies. I thank Dr. Prathiba Natesan Batley for serving on a whim and for providing insightful feedback. Dr. Dena Dossett, thank you for your thoughtful and gracious JCPS leadership and for supporting my dissertation and professional growth in countless ways.

My early graduate school days were brighter because of the kindness of Tammy Green and the friendship of Alireza Aghaey, Chuck Olsavsky, Tommie Welcher, and the incomparable Cole Crider. I am grateful for superb scholarly guidance from Drs. Jim Fiet, Manju Ahuja, Namok Choi, Jill Adelson, Kate Snyder, Amanda Mitchell, and Kyle Ingle. I thank Tamara Lewis, Leslie Taylor, Patrick Cyrus, Ryan McCafferty, Thomas Reece, and Ben Wilborn of JCPS for supporting my CSS research. Last, but certainly not least, working with the amazing team of Joe Prather, Florence Chang, Fiona Hollands, Rob Shand, and the legendary Bo Yan has truly been a life-changing opportunity to grow as a researcher, practitioner, and person. *Nanos gigantum humeris insidentes*.

ABSTRACT

A PSYCHOMETRIC ANALYSIS OF THE COMPREHENSIVE SCHOOL SURVEY (CSS) MIDDLE SCHOOL STUDENT VERSION

By

Stephen M. Leach

November 11, 2022

School climate is increasingly recognized by scholars and policymakers as a crucial factor associated with students' educational experiences. Hence, practitioners endeavor to equitably measure and improve school climate to promote favorable student academic and behavior outcomes. Unfortunately, school climate research is fragmented, and a research-practice gap exists in best scale development and validity testing practices. The result is a proliferation of practitioner-developed school climate measures lacking solid theory-grounding and evidence to support intended score interpretations and uses.

In response, Whitehouse et al. (2021) proposed a validity testing framework for practitioner-developed instruments aimed at supporting culturally responsive school climate measurement. Their framework, however, suffers from key limitations regarding the transparency of content validity assessment, breadth of validity evidence reported, and methods used to examine measurement invariance. Therefore, this study sought to replicate and extend their validity testing framework by using a standardized rubric to assess content validity, examining measurement invariance via the alignment method, and analyzing the predictive validity of group mean scores. By applying the extended

validity testing framework to a practitioner-developed school climate student survey, the study also aimed to provide useful evidence to a large urban district with respect to the validity of comparing survey scores across Black and White middle school student groups and using scores to inform continuous improvement of student learning

Results suggest the extended framework is superior to the original for obtaining general content, factorial, and predictive validity evidence, and assessing measurement invariance across racial subgroups, provided the number and size of groups are adequate. Findings suggest the district's middle school student survey is culturally responsive, although it may not sufficiently address all critical school climate dimensions. To improve the survey, the district must settle on a clear definition and taxonomy of school climate to facilitate a program of validity testing, and publicly document all available validity evidence. Future studies should clarify alignment sample size and simulation study requirements and extend the framework to assess additional validity concerns and for use with person-centered approaches.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

INTRODUCTION

School climate is a crucial factor associated with students' educational experiences (e.g., Jordan & Hamilton, 2020; Ryberg et al., 2020, Rudasill et al., 2018; Wang & Degol, 2016). Although the construct's precise definition and dimensional structure remain unsettled, a positive school climate is said to be one in which a school's community (e.g., students, staff, and parents) feels safe, supported, and engaged while working together towards a shared vision that includes placing a high value on education (Cohen et al., 2009). School climate, therefore, represents a combination of subjective judgements about a school's environment (e.g., Bradshaw et al., 2021; Cohen & Thapa, 2017; Rudasill et al., 2018). The U.S. Department of Education (ED, 2014), National Education Association (NEA; e.g., Long, 2017), and National School Climate Center (NSCC, 2021) have advocated for improving school climate as a means of promoting more favorable academic and behavioral student outcomes. A key assumption underlying the emphasis on school climate improvement is that the perceptions of students, teachers, parents, and other key stakeholders can be reliably measured, and obtained scores can be interpreted and used for decision-making purposes (e.g., Clifford et al., 2012; Ryberg et al., 2020; Schweig et al., 2019).

This assumption regarding measurement is evident in the proliferation of school climate surveys developed by researchers and national-, state-, and local education agencies (e.g., Bear, Yang, Mantz et al., 2014; ED, 2020; Hough et al., 2017; Huang et al., 2015; Lewis, 2019). Indeed, self-report surveys are by far the primary method used to

measure school climate (e.g., Bradshaw et al., 2021; Schweig et al., 2019; Wang & Degol, 2016). Conceptually, school climate is considered a multidimensional construct (e.g., Rudasill et al., 2018) comprised of such domains as safety and interpersonal relationships (e.g., Bradshaw et al., 2021; Lewno-Dumdie, 2020; Wang & Degol, 2016). Regrettably, however, many school climate instruments lack sufficient reliability and/or validity evidence to support intended score interpretations and uses (e.g., Bear et al., 2015; Cohen et al., 2009; Cohen & Thapa, 2017; Ramelow et al., 2015). This problem is especially prevalent among practitioner-developed instruments (Bear et al., 2015; Ramelow et al., 2015), largely because the validity literature is unfamiliar or inaccessible to practitioners (Schweig et al., 2019). Without supporting evidence, well-intentioned educators may make ill-founded decisions based on; (a) improper subscale scores (Leach et al., 2020), (b) uninterpretable group score differences (Asparouhov & Muthén, 2014), and/or (c) unknown relationships with other outcomes (Schweig et al., 2019).

To be fair, researchers have conducted validity studies on a number of school climate measures (e.g., Bear et al., 2011, 2015, 2018; Ryberg et al., 2020; Whitehouse et al., 2021). As a result, several existing school climate measures have at least some evidence supporting the validity of their intended score uses and interpretations (c.f., American Educational Research Association [AERA], et al., 2014; Kane, 2013; Messick 1989, 1995). Unfortunately, existing instruments frequently do not meet educators' practical needs (Bear et al., 2015; Clifford et al., 2012) in terms of target population(s), survey length, and/or cost. School climate surveys must assess all preferred respondent types (Bear et al., 2015) and must not place undue burdens on respondents' time or administrators' budgets (Clifford et al., 2012; Waasdorp et al., 2019). Thus, an important

question for practitioners interested in measuring and improving school climate is whether to select one of the many existing scales or to create their own (e.g., Cohen & Thapa, 2017). When educators elect to develop an instrument, the dual practicality-reliability/validity requirement suggests a need for researcher-practitioner partnerships to assess reliability and validity (Whitehouse et al., 2021).

Acknowledging the ubiquity of locally developed school climate measures that lack a clear theoretical basis (e.g., Ramelow et al., 2015) and the need for culturally responsive instruments (e.g., Bear et al., 2011; Zabek et al., 2022) in urban school settings, Whitehouse et al. (2021) proposed a collaborative two-part validity testing framework for such measures. They recommend first assessing construct validity by (a) ascertaining the instrument's factor structure (i.e., factorial validity), and (b) assessing item representativeness and domain coverage (i.e., content validity) based on correspondence of the scale's factors with an existing school climate model. Whitehouse et al. used five school climate dimensions proposed by Thapa et al. (2013) to determine content validity. Their second step is to test the preferred model for measurement invariance (MI) across racial/ethnic subgroups of interest to facilitate, if MI is exhibited, analyzing between-group score differences.

Whitehouse et al. (2021) have therefore taken an important first step in proposing a standardized validity testing framework for practitioner-developed school climate measures. Their framework, however, suffers from key limitations with respect to the transparency of their content validity assessment process, the breadth of validity evidence reported, and the use of multi-group confirmatory factory analysis (MGCFA) to test MI. Besides limiting comparisons to two or three groups, the application of MGCFA to

Likert-type items is tenuous (Flake & McCoach, 2018) and it rarely produces strict invariance (Marsh et al., 2018). To address these issues, Asparouhov and Muthén (2014) developed the alignment method. Flake and McCoach applied this method to polytomous items in a simulation study and found that the approach acceptably recovered parameter estimates, permitting group comparisons even in the presence of moderate noninvariance.

The primary purpose of this study is to replicate and extend Whitehouse et al.'s (2021) framework in terms of the transparency of the validity process, the statistical techniques employed, and the breadth of validity evidence reporting. To accomplish that task, I investigated a school climate measure developed and routinely administered by a large, urban public school district. The second aim of this study is to provide useful evidence to the school district regarding the validity of comparing school climate survey scores across middle school student racial/ethnic subgroups and using scores to inform school improvement efforts. The updated framework has broad applications beyond the specific score uses and interpretations investigated here. The validity process described below can guide the district in assessing the validity of alternative score interpretations (e.g., analyzing long-term school climate trends) and uses across all versions of the scale (e.g., parent and staff). The framework is not limited to validity studies of school climate in urban school contexts but can be adapted for a variety of constructs and settings where instruments were developed apart from best practices.

LITERATURE REVIEW

According to Kohl et al. (2013), educators seeking to measure school climate can take one of three basic instrumentation approaches: adoption, adaption, or creation. They describe an ideal scenario in which existing scales with acceptable reliability and validity evidence are first examined systematically to determine whether they can be adopted or adapted for a particular use and chosen conception of school climate. Reliability refers to the degree to which an instrument's scores are consistent and without measurement error (e.g., Schweig et al., 2019); validity refers to the degree of support for score uses and interpretations provided by theory and evidence (AERA et al., 2014).

Per the *Standards for Educational and Psychological Testing* (hereafter *Standards*; AERA et al., 2014), validity should be the primary concern in instrument development. In practice, however, educators often skip straight to creation without a clearly delineated approach to school climate (e.g., Ramelow et al., 2015; Schweig et al., 2019; Whitehouse et al., 2021). The lack of theory-grounding in school climate measures gives little guidance for assessing content validity, or the degree to which items align with theorized dimensions (Schweig et al., 2019), and often results in incomplete domain coverage (Ramelow et al., 2015).

One example of an educator-created scale is Jefferson County Public Schools' (JCPS) *Comprehensive School Survey* (CSS; JCPS, 2018a; Lewis, 2019; Muñoz & Lewis, 2009; Rudasill & Rakes, 2008). JCPS, a large, urban school district in Louisville, KY, has administered the CSS annually to students, parents, and employees since the

1996-97 academic year. The district provides several publicly accessible online tools for examining CSS results (JCPS, 2018a) and encourages teachers and school administrators to make use of CSS results (JCPS, 2018b). Furthermore, JCPS leadership has indicated that CSS scores will inform both improvement efforts (Tatman, 2018) and strategic planning (JCPS, 2018a). The emphasis on using CSS results is not surprising, given that *Climate & Culture* is one of JCPS' three key pillars for improving learning (JCPS, 2018c). Indeed, the district maintains a School Climate and Culture Department which, among other things, oversees multi-tiered systems of support and social-emotional learning efforts (JCPS, 2018d).

What is surprising, however, is that despite the sharp focus on improving school climate (JCPS, 2018d; Tatman, 2019), JCPS has no formal definition of school climate. In publicly available online documentation, the district appears to use the terms *climate* and *culture* interchangeably (e.g., Tatman, 2018), implying that whether used separately or together, they represent a single latent construct (c.f., Rudasill et al., 2018). Some documentation suggests that JCPS views school climate as a higher-order construct comprised of 14 subdimensions, referred to as *constructs* (JCPS, 2019a). The theoretical and/or empirical bases for the underlying factor structure are unclear based on publicly available information (c.f. *Standards*, Standard 1.13).

The lack of conceptual clarity can also be seen in conflicting construct names and item compositions between the various outputs obtained by online CSS comparison tools (JCPS, 2018a). For example, the item "I feel safe on my way to and from school" is listed as one of three items comprising a *Personal Safety* subscale but also included in the 18-item *School* subscale. The district's CSS Results tool reports item-level scores for the

School subscale whereas the CSS Constructs tool reports a single, undescribed score (perhaps percent agreed?) for the *Personal Safety* subscale. Leach et al. (2020) suggested that subscale scores may not be appropriate for instruments with complex factor structures comprised of intentionally multidimensional items. Item level scores should not be compared without evidence of residual invariance (Saint et al., 2021). In a 2019 report, the district provided an index of 13 items which form five subdimensions that are referred to as "constructs related to culture and climate" (JCPS, 2019b). Although the processes are not well-documented (c.f. *Standards,* Standard 1.11), the committee-based approach to CSS development (e.g., Lewis, 2019) indicates that JCPS has at least partially followed the recommendation of Olsen and colleagues (2017) to consider both the composition and indicators of the CSS 'culture and climate' construct. However, the lack of a well-defined underlying conception of school climate (e.g., Ramelow et al., 2015) and the conflicting subdimension and item selection issues outlined above call into question both what the CSS is intended to measure and what it actually measures.

Acknowledging the apparent research-practice gap exemplified by, but certainly not limited to, the CSS, Whitehouse et al. (2021) proposed a validity testing framework for practitioner-developed school climate measures lacking an explicit theoretical underpinning. In essence, their framework is a means of recreating (or entering) the ideal adoption-adaption-creation sequence described by Kohl et al. (2013). In the Whitehouse et al. approach, the sequence is preceded by gathering construct validity evidence, namely evidence of content (e.g., *Standards*, Standard 1.9, 1.11) and factorial, or internal structure validity (e.g., *Standards*, Standard 1.13-1.15), for the educator's existing instrument. Based on the strength of construct validity evidence obtained via the

Whitehouse et al. framework, educators may choose to adopt (i.e., continue using) or adapt (i.e., revise based on newly obtained validity evidence) their own measure. If the evidence does not support the intended uses of their measure, practitioners may begin the adoption-adaption-creation process afresh by examining whether other existing measures fit their needs. Crucially, the first step in any of the ideal or real-world measurement scenarios just described is the same; educators must begin with a clear definition of school climate (e.g., Chirkina & Khavenson, 2018; Kohl et al., 2013; Lewno-Dumdie et al., 2019; Olsen et al, 2017; Schweig et al., 2019; Whitehouse et al., 2021).

Defining School Climate

Lack of Consensus

Unfortunately, perhaps the only consensus around defining school climate is that there is no consensus (e.g., Berkowitz et al., 2017; Cohen et al., 2009; Cornell et al., 2017; Huang & Cornell, 2016; Kohl et al., 2013; Lewno-Dumdie et al., 2019; Lindstrom Johnson et al., 2019; Ramelow et al., 2015; Rudasill et al., 2018; Ryberg et al., 2020; Schweig et al., 2019; Shukla et al., 2019; Thapa et al., 2013; Wang & Degol, 2016; Whitehouse et al., 2021). Grazia and Molinari (2022) and Berkowitz et al. (2017) adjure researchers to settle on a universal definition to alleviate confusion and facilitate adequate measurement to support long-term school improvement, although at present that solution seems unlikely. Rudasill et al. (2018) suggest that the confusion between various conceptions of school climate stems in part from the failure of researchers to distinguish between definitions, taxonomies, and models. For them, a precise, operational definition should provide a clear boundary line as to what is and is not school climate and the corresponding taxonomy should categorize the dimensional structure of the underlying

causal model, or theory. The underlying model outlines theorized relationships between school climate dimensions, the hypothesized mechanisms through which school climate is formed, and proposed associations between school climate dimensions and other outcomes (Rudasill et al., 2018). Wang and Degol (2016) also differentiate between concrete (i.e., operational) and abstract (i.e., conceptual) definitions and call on researchers to better delineate the school climate taxonomy.

Common Ground

The lack of agreement on an operational definition of school climate has clear measurement implications which will be discussed below. The state of disagreement, however, does not imply a complete lack of commonality among many of the competing school climate definitions (e.g., Rudasill et al., 2018). Scholars and practitioners generally agree (e.g., Grazia & Molinari, 2022; Olsen et al., 2017) with the NSCC's (2021) broad depiction of school climate as referring to the quality and character of school life. Researchers also largely agree that school climate is a malleable and complex, multidimensional school-level construct comprised of the aggregated perceptions of various members of a school's community regarding specific aspects of school life (e.g., Cohen et al., 2009; Grazia & Molinari, 2022; Rudasill et al., 2018; Thapa et al., 2013; Wang & Degol, 2016; Zullig et al., 2015). Seemingly all extant conceptual models (e.g., Aldridge & McChesney, 2021; Bear, Yang, Mantz, et al., 2014; Bradshaw et al., 2021; Cohen & Thapa, 2017; Hough et al., 2017; Kohl et al., 2013; Konold et al., 2021; Lewno-Dumdie et al., 2020; NSCC, 2021; Rudasill et al., 2018) include students and school staff as pertinent school community members whose combined experiences of school life constitute a school's climate. Parent perceptions

(e.g., Bear et al., 2011, 2015, 2018) of school life partly compose school climate in many, though not all, models (e.g., Rudasill et al., 2018).

Furthermore, several reviews have suggested common domains of school climate across extant studies. For example, Cohen and colleagues (2009) identified safety, teaching and learning, relationships, and environmental/structural domains, to which they later added school improvement processes (Thapa et al., 2013). Wang and Degol (2016) found academic climate, community, safety, and institutional environment domains evident among 327 empirical and conceptual studies they reviewed. Rudasill et al. (2018) identified shared beliefs and values, relationships and social interactions, safety, teaching and instruction, leadership, and physical environment as consistent themes emerging from research in the organizational, school effects, and psychology literatures. Thus, at least some school climate domain consistency (e.g., safety, relational, and institutional factors) exists across time, theoretical approaches, and research traditions.

NSCC Definition

The foregoing discussion implies that the NSCC's (2021) definition of school climate as the quality and character of school life is incomplete. The broad, abstract nature (e.g., Wang & Degol, 2016) of the definition may explain why it is so widely cited (e.g., Bradshaw et al., 2021; Cohen et al, 2009; Lewno-Dumdie et al., 2019; Marx & Byrnes, 2012; Olsen et al., 2017; Ramelow et al., 2015; Thapa et al., 2013; Zullig et al., 2010; 2015), even among studies that do not advocate the same domains and dimensions proposed by the NSSC (e.g., Aldridge & Ala'l, 2013; Bear et al, 2011; Kohl et al., 2013; Konold & Cornell, 2015; Konold et al., 2021; Lindstrom Johnson et al., 2019; Wang & Degol, 2016). Upon closer inspection, however, the characterization of the NSCC's

definition of school climate as abstract and incomplete appears to overstate the case. This is because the Center does not *only* define school climate in those terms (e.g., Berkowitz et al., 2017) but goes on to delineate the construct more concretely (e.g., Rudasill et al., 2018; Wang & Degol, 2016) and provide a taxonomy (e.g., Rudasill et al., 2018).

What School Climate Is. In the NSCC's (2021) estimation, school climate is comprised of the aggregated perceptions of a school community (i.e., students, parents, and school personnel) along five domains: *safety*, *teaching and learning*, *interpersonal relationships*, *institutional environment*, and *social media.* As such, school climate reflects a school's norms, goals, values, practices, and organizational structures.

What School Climate Is Not. Because school climate is based on the experiences of school community members, neither aggregated nor disaggregated objective measures of each of the five domains or demographic characteristics of the school community are included (e.g., Berkowitz et al., 2017; Bradshaw et al., 2021; Kohl et al., 2013; Rudasill et al., 2018). Similarly, ratings of self-efficacy are not considered a component of school climate (e.g., Hough et al., 2017). Ratings of school administrators are also excluded (c.f., Cohen et al., 2009) from the NSCC's five domains in its survey for students and parents.

Dimensions of School Climate

This section follows the example of Rudasill et al. (2018) by presenting a taxonomy of the school climate model proposed by the NSCC (2020). Here, we consider the dimensional structure of the model in greater detail, drawing heavily from the NSCC's website (www.schoolclimate.org). I will also briefly discuss each dimension's alignment (or not) with other school climate models. For clarity, the discussion is organized by the NSCC's (2020) five school climate domains outlined above.

Safety. Three dimensions–*Rules and Norms*, *Physical Security*, and *Social–Emotional Security*–are included in the safety domain. *Rules and Norms* explicitly convey expectations and consequences regarding violence, harassment, and verbal abuse. *Physical* and *Social-Emotional Security* indicate perceived safety from physical violence forms of verbal abuse (including exclusion), respectively.

These dimensions of school safety are ubiquitous (e.g., Bradshaw et al., 2021; Lewno-Dumdie, 2020; Wang & Degol, 2016) among existing models although all three are not always included within the safety domain, or at all, in each taxonomy. For example, physical safety, emotional safety, and bullying dimensions partly comprise the U.S. Department of Education's (ED; e.g., Ryberg et al., 2020) safety domain, although ED considers elements of rules and norms as belonging to an institutional environment domain. Much psychometric support has been found for the *Delaware School Climate Survey* model (DSCS; e.g., Bear et al., 2011, 2015, 2018; Bear, Yang, Mantz, et al., 2014; Bear, Yang, Pell et al., 2014; Yang et al., 2013, 2021), which currently proposes clarity of expectations, fairness of rules, bullying, and school safety as four distinct dimensions of school climate within the Demandingness and Structure domain (Bear, Yang, Mantz, et al., 2014). The precise dimensional structure of the components included in some taxonomies appears to be empirically derived rather than conceptually delineated (e.g., Bear, Yang, Mantz, et al., 2014; Whitehouse et al., 2021). When the underlying dimensionality indicated by theory and factor analysis are misaligned, uncertainty exists about the need for new theorizing or better measurement, or both, to remedy the misalignment (e.g., Leach et al., 2020).

Interpersonal Relationships. *Respect for Diversity* entails a sense of appreciation for distinctive individual characteristics and expectations of tolerance among school community members. Two additional dimensions–*Social Support-Adults* and *Social Support-Students*–express the extent to which students' relationships with the adults and students in their school community are characterized by high expectations for learning, personal concern for problems, and a sense of welcoming for new students.

As with safety, there is near universal agreement (c.f., Rebelez & Furlong, 2013) among scholars and practitioners that healthy relationships between school members are a critical element of a positive school climate (e.g., Berkowitz, 2017; Bradshaw et al., 2021; Cohen et al., 2009; Grazia & Molinari, 2022; Lewno-Dumdie et al., 2020; Olsen et al., 2017; Rudasill et al., 2018; Thapa et al., 2013; Wang & Degol, 2016). However, just as with safety, precise dimensional structures vary across models. For example, the DSCS model (e.g., Bear, Yang, Mantz, et al., 2014) includes similar components as the NSCC whereas ED (2020) considers student-student and student-teacher relationships as comprising a single dimension in the engagement domain. Within that same domain, a cultural and linguistic competence dimension separately addresses elements of diversity (e.g., Ryberg et al., 2020). Interestingly, the five dimensions proposed by Whitehouse et al. (2021) do not include a separate relationship dimension, although elements of student-student and teacher-student relationships are included in the other dimensions. Their results are largely empirically driven, although that is not surprising given their intent to find conceptual support for an existing locally developed scale.

Teaching and Learning. *Support for Learning* includes experiencing supportive, differentiated teaching practices aimed at fostering independent thinking, dialogue, and

manifold avenues for demonstrating skill mastery. *Social and Civic Learning* reflects perceived encouragement towards social and civic awareness and engagement, with emphasis on effective communication and successfully making ethical decisions and navigating conflicts. Although Cohen and colleagues (Cohen et al., 2009; Cohen, 2013, 2017; Thapa et al., 2013) contend that these dimensions are an essential aspect of school climate, Rudasill et al. (2018) argue that they are not components but instead influence school climate by affecting the formation of interpersonal relationships.

In contrast to Rudasill et al. (2018), many school climate models include aspects of teaching and learning (Grazia & Molinari, 2022; see Rebelez & Furlong, 2013 and Bear, Yang, Mantz et al., 2014 for exceptions). Lewno-Dumdie et al. (2020) found that 13 of the 18 measures they investigated included dimensions of teaching and learning. Wang and Degol's (2016) school climate review categorized teaching and learning as a dimension of academic climate. Compared with the NSCC's safety and interpersonal relationships domains, teaching and learning appears to be less universally regarded by scholars as a crucial component of school climate (e.g., Bear et al., 2011, 2015, 2018; Bear, Yang, Mantz et al., 2014). Even among models that include aspects of teaching and learning, researchers disagree whether it is a domain (e.g., NSCC, 2020) or dimension (e.g., ED, 2020) or subdimension (e.g., Sun & Royal, 2017).

Institutional Environment. The three dimensions of this domain cover the sense of positive association with a school's customs and traditions, and involvement in the many aspects of school life (*School Connectedness/Engagement*) and include perceptions about how the school community seeks to welcome, affirm, and involve its members with disabilities (*Social Inclusion*). The final dimension, *Physical Surroundings*, encompasses

the perceived condition of a school's physical environment, including the availability and sufficiency of materials and resources.

Support for these dimensions among existing school climate models is mixed. For example, some researchers do not include physical surroundings at all in their taxonomies of school climate (e.g., Bear, Yang, Mantz, et al. 2014; Rebelez & Furlong, 2013; Rudasill et al., 2018; Sun & Royal, 2017). Grazia and Molinari (2022) found physical and resource availability dimensions present in only a small number of studies they reviewed (c.f., Ryberg et al., 2020; Saint et al., 2021). Lewno-Dumdie (2020) found more than 80% of the instruments they reviewed measured at least some dimensions included in the NSCC's (2020) Institutional Environment domain, although less than half assessed physical surroundings. Aspects of student connectedness and engagement are especially common among extant models (e.g., Grazia & Molinari, 2022), although many researchers do not consider connectedness to be a dimension of a school's environment (e.g., Bear, Yang, Mantz et al., 2014; Rudasill et al., 2018; Ryberg et al., 2020; Saint et al., 2021; Wang & Degol, 2016, You et al., 2014). Aspects of inclusion specific to students with disabilities are noticeably absent from most current models (e.g., Bear, Yang, Mantz et al., 2014; ED, 2020; Saint et al., 2021) as evidenced by a lack of mention in several recent school climate reviews (e.g., Bradshaw et al., 2021; Chirkina & Khavenson, 2018; Grazia & Molinari, 2022; Lewno-Dumdie et al., 2020; Rudasill et al., 2018; Wang & Degol, 2016).

Social Media. Consisting of a single dimension, social media indicates a feeling of safety from harm (e.g., teasing, exclusion, verbal abuse) when students are online. Although some researchers view school climate and bullying, including cyberbullying, as

interrelated constructs (e.g., Grazia & Molinari, 2022), perceptions of experiencing these forms of harm are often included in school climate taxonomies (c.f., You et al., 2014) as individual factors (e.g., Konold & Cornell, 2015) or as dimensions of the safety domain (e.g., Bear, Yang, Mantz, et al., 2014; Berkowitz et al., 2017; Bradshaw et al., 2021; ED, 2020; Rudasill et al., 2018; Wang & Degol, 2016). Indeed, even the NSCC's (2020) taxonomy includes a sense of physical and social-emotional safety under its Safety domain, thus it is unclear why online forms of these behaviors should constitute a separate domain altogether.

Rudasill et al. (2018) suggest that much of the incertitude regarding school climate research stems from conflating definitions, taxonomies, and models. Concretely defining school climate per the NSCC and outlining its corresponding taxonomy has enabled us to examine the alignment (or not) between the NSCC's conceptualization of school climate and competing definitions. Having addressed two of the three key elements of school climate confusion identified by Rudasill and colleagues, I now turn to the NSCC's (2021) school climate model.

Associated Outcomes

According to the National School Climate Council (2007), the patterns of judgements formed by various groups of a school's community regarding safety, relationships, engagement, emphasis on academic, social, and civic learning, and the physical environment either foster or inhibit an effective learning environment, i.e., a positive school climate. A positive climate, in turn, is hypothesized to predict favorable student academic and behavioral outcomes (e.g., Cohen et al., 2009; Daily et al., 2019; ED, 2014; Long, 2017; Reaves et al., 2018; Schweig et al., 2019; Thapa et al., 2013;

Wang & Degol, 2016; Zullig et al., 2011). Researchers have found positive relationships between school climate and self-reported academic achievement (Daily et al., 2019), grades (Hopson & Lee, 2011), and standardized test scores (MacNeil et al., 2009) among middle and high school students. Berkowitz and colleagues' (2017) synthesis of 78 studies published after the year 2000 suggested that school climate may mediate the negative relationship between socioeconomic status and academic achievement. Gage et al. (2016) found a negative relationship between school climate and discipline referrals among a sample of K-12 students. Huang and Cornell (2018) reported a similar relationship between school climate and out-of-school suspensions for middle school students. Steffgen et al. (2013) meta-analyzed 36 empirical studies and reported a moderate negative effect (r = -.26) of school climate on school violence.

The NSCC's (2021) current definition and taxonomy are based largely on syntheses (e.g., Cohen et al., 2009; Thapa et al., 2013) of empirical school climate research conducted prior to 2012, which the center organizes under its five domains of school climate (i.e., safety, teaching and learning, institutional environment, interpersonal relationships, and social media). The Center's classification scheme for extant research, however, is not well-described and at times contradictory. For example, a study by Lee et al. (2011) on the relationship between school suspensions and dropout rates is included in both the safety and teaching and learning domains. However, Catalano et al.'s (2004) research on the relationship between school connectedness and risky behaviors is placed in the safety domain but not in the institutional environment domain despite school connectedness being included in the latter in the NSCC's taxonomy. The ongoing lack of consensus over definitions and taxonomies of school climate (e.g., Rudasill et al., 2018;

Wang & Degol, 2016) and unresolved issues in school climate measurement, which are discussed in greater detail below, undoubtedly contribute to the lack of clarity in the NSCC's classification scheme for existing school climate research. Because nearly all empirical school climate studies are correlational (e.g., Berkowitz et al., 2017; Bradshaw et al., 2021), the choice of classifying studies based on school climate dimensions as predictors versus outcomes is somewhat arbitrary. Taken together, the existing research suggests a complex relationship between school climate dimensions and academic and behavior outcomes and highlights the need for psychometrically sound school climate measures to help future research further our understanding of these relationships.

The preceding discussion highlights what Rudasill et al. (2018, p. 41) refer to as the 'chaotic conceptual landscape' of school climate research. On the one hand, the NSCC's (2020) broad definition of school climate as the quality and character of school life is widely cited in the literature and its proposed dimensions are common among alternative school climate models (e.g., Lewno-Dumdie et al., 2020). On the other hand, concretely defining school climate in the NSCC's terms and more clearly delineating the organization's taxonomy exposes incongruences in domain coverage and dimensional structures across existing models and, in the case of social media, seemingly within the NSCC's model itself. With this in mind, let us now turn to school climate measurement.

Measuring School Climate

Self-Report Surveys

The current state of school climate measurement reflects the overlapping, yet unsettled nature of school climate definitions, taxonomies, and models (e.g., Bradshaw et al., 2021; Cohen & Thapa, 2017; Grazia & Molinari, 2022; Huang et al, 2015; Konold et

al., 2021; Lewno-Dumdie et al., 2020; Rudasill et al., 2018). The inability of scholars to agree on a unified model of school climate (e.g., Bradshaw et al., 2021) is clearly seen in the proliferation of competing instruments designed to measure various conceptions of school climate (Grazia & Molinari, 2022; Kohl et al., 2013; Lewno-Dumdie et al., 2020; Rudasill et al., 2018). Despite numerous instruments resulting from the fragmented nature of school climate theory (Grazia & Molinari, 2022), scholars have found commonalities among the measurement approaches taken in published school climate studies (e.g., Bradshaw et al., 2021; Kohl et al., 2013; Lewno-Dumdie, 2020; Wang & Degol, 2016; Zullig et al., 2010).

Foremost among shared elements in school climate measurement is reliance on self-report surveys. Wang and Degol (2016) estimate more than 90% of published empirical school climate studies used self-report surveys, most with Likert-type items. Lenz et al. (2021) reviewed nine school-climate surveys developed between 1993 and 2017 and found item counts ranging from nine to 153, although the majority (six) were comprised of between 29 and 54 items. School climate surveys are most commonly administered to students (e.g., Bear et al., 2011; Ryberg et al. 2020; Zullig et al., 2015), school staff (e.g., Bear, Yang, Pell et al., 2014), and parents/guardians (e.g., Bear, 2015), respectively.

Given dimensional overlaps among many competing school climate models (e.g., Rudasill et al., 2018), instruments based on alternate conceptual models unsurprisingly tend to measure distinct yet overlapping dimensions (e.g., Lewno-Dumdie et al., 2020; Olsen et al., 2017). For example, all 18 school climate measures reviewed by Lewno-Dumdie and colleagues (2020) included various combinations of at least three of the five

dimensions (safety, relationships, teaching and learning, institutional environment, and school improvement processes) advocated by Cohen et al. (2009) and Thapa et al. (2013). Lewno-Dumdie et al. interpreted their findings to indicate a weak consensus on dimensionality.

Reliability and Validity Reporting

Inadequate or unavailable score reliability and validity information is prevalent among extant school climate measures (e.g., Bear et al., 2015; Cohen et al., 2009; Olsen et al., 2017; Ramelow et al., 2015; Zabek et al., 2022). Of 26 surveys reviewed by Olsen and colleagues (2017), only four met their acceptable technical adequacy (i.e., reliability and validity reporting) criteria. Among those four instruments that did report reliability, internal consistency reliability coefficients were below .70 for several subscales and no validity details were provided. Although Ramelow et al. (2015) found acceptable reliability (i.e., Cronbach's alpha $\geq .70$) among eight of 12 published school climate scales, they found validity evidence largely inadequate or omitted altogether.

Despite Messick's (1989, 1995) admonition that validity is a property of specific score interpretations and uses, school climate researchers frequently refer to instruments themselves as valid (e.g., Aldridge & Ala'l, 2013; Aldridge & McChesney, 2020; Kohl et al., 2013, Olsen et al., 2017; Whitehouse et al., 2021). Attributing validity to an instrument wrongly implies that its scores can be validly interpreted and used for any purpose (c.f., AERA et al., 2014; Clifford et al., 2012; Schweig et al., 2019). Kohl et al. (2013) suggest educators first consider adopting or adapting an existing school climate scale with adequate validity evidence, however, the widespread lack of reliability and validity reporting among published scales limits the available choices.

The Practitioner World

To develop school climate scales that meet the dual requirements of practicality and reliability/validity, educators must be aware of the need – and be willing and able – to formally investigate the psychometric properties of scores derived from employed school climate measures (c.f., Schweig et al., 2019). In the 'real world' of K-12 education, practitioners often skip crucial steps when creating instruments (e.g., Cohen & Thapa, 2017; Hamilton et al., 2019; Zullig et al., 2010). Locally developed school climate surveys frequently lack a clear, underlying theoretical basis, resulting in uneven domain coverage across instruments (e.g., Ramelow et al., 2015). Educators are not typically trained in psychometrics and measurement and, therefore, likely unaware (e.g., Cohen & Thapa, 2017; Schweig et al., 2019) of both the science of survey development (e.g., de Leeuw et al., 2014) and the need to gather specific types of evidence to support scoring inferences (e.g., AERA et al., 2014; Kane, 2013; Messick, 1989; 1995; Schweig et al., 2019).

When schools and districts create and administer school climate surveys apart from established best-practices (e.g., AERA et al., 2014; de Leeuw et al., 2014, Difazio et al., 2018), the resulting response scores may not be reliable (e.g., Henson, 2001; Ramelow et al., 2015) and, to the extent that agencies fail to gather validity evidence, score interpretations and uses may be called into question (e.g., AERA et al., 2014; Kane, 2013; Messick, 1989, 1995; Whitehouse et al., 2021). In such commonplace scenarios, (e.g., Hamilton et al., 2019; Ramelow et al., 2015; Schweig et al., 2019; Zullig et al., 2010), practitioners are essentially leaving the appropriateness of their interpretations and uses of survey scores to chance.

The general consensus is that school climate is worth improving (e.g., Bradshaw et al., 2021; Cohen et al., 2009; ED, 2014; Long, 2017; NSCC, 2021; Ramelow et al., 2015; Ryberg et al., 2020; Rudasill et al., 2018; Wang & Degol, 2016; Whitehouse et al., 2021). If then, school climate measurement is too important to be left to chance (e.g., Ramelow et al., 2015), does this mean that educators should cease administering their own existing school climate instruments and begin the instrument development process afresh, with a firm and explicit grounding in theory (e.g., Cohen et al., 2009; Ramelow et al., 2015; Van Houtte & Van Maele, 2011) and guided by the latest methodology (e.g., AERA et al., 2014; de Leeuw at al., 2014; Finch et al., 2016; Difazio et al., 2018)? While that is an option, a practically feasible, science-based alternative is to partner with a research team to gather validity evidence supporting the survey's intended interpretations and uses (e.g., Bear et al., 2011, 2015, 2018; Kane, 2013; Messick, 1989, 1995; Ramelow et al., 2015; Ryberg et al., 2020; Schweig et al., 2019; Whitehouse et al., 2021; Zullig et al., 2015).

Validity Testing Framework

A psychometric investigation of a locally developed school climate survey can provide leadership with evidence-based confidence in response score interpretations and their intended uses where warranted (e.g., AERA et al. 2014; Kane, 2013; Messick, 1989, 1995), and guidance for any needed improvements (e.g., de Leeuw et al., 2014; Finch et al., 2016; Marsh et al., 2018). Indeed, Hollands et al. (2022) recommend this approach for JCPS *Comprehensive School Survey*. Whitehouse et al. (2021) recently investigated a locally developed school climate measure lacking a clear theoretical underpinning. In doing so, the authors provide a validity-testing framework for similar research projects in

urban ‘majority-minority’ districts where educators need culturally responsive measures (e.g., Zabek et al., 2022) to inform climate improvement efforts. Whitehouse and colleagues sought to obtain validity evidence to support comparisons of school climate survey scores across racial/ethnic subgroups. Their framework is applicable to the current study, given the makeup of JCPS’ student population and the urban district’s focus on equitably improving climate.

Whitehouse et al. (2021) followed traditional factor analytic approaches. First, they conducted exploratory, followed by confirmatory factor analysis (EFA-CFA) on separate samples (e.g., Morin et al., 2013). To assess construct representation (e.g., *Standards*, Standard 1.9, 1.11), they compared the resulting five factors to Thapa and colleagues’ (2013) proposed school climate dimensions (c.f., Zullig et al. 2014). Second, the authors conducted multi-group confirmatory factor analysis (MGCFA) to test progressively more restrictive models for configural, metric, scalar, and strict invariance (e.g., Flake & McCoach, 2018). Finally, despite failing to achieve strict invariance (c.f., Asparouhov & Muthén, 2014; Flake & McCoach, 2018; Marsh et al., 2018) the authors analyzed standardized factor score differences between three racial/ethnic subgroups using ANOVA. Whitehouse et al. (2021) have sought to fill a critical research-practice gap by proposing a standardized validity testing framework for locally developed urban school climate measures. However, their approach suffers from key limitations regarding the breadth of reliability and validity evidence gathered and/or reported, and the methodological approach to examining MI.

Types of Reliability and Validity Evidence

Per the *Standards* (Standard 1.1–1.5), scale development entails identifying proposed score uses and/or interpretations and providing either supporting evidence or relevant disclaimers if no validity evidence is available. In the latter case, which is in focus here, the disclaimer is meant to temporarily warn users about potential misuses and/or misinterpretations while validity evidence is being gathered (*Standards*, Standard 1.3, 1.4). To provide researchers with practical guidance for assessing validity, scholars have advanced the argument-based approach to validity (e.g., Chapelle et al., 2008; Cronbach, 1988; Kane, 2006, 2013; Shepard, 1993), which, in simplest terms, proposes that researchers gather and evaluate only that validity evidence which pertains to intended score interpretations and uses. An argument for specific score interpretation and uses is clearly delineated and then evidence to support that argument is gathered and evaluated (e.g., Kane 2013). The argument-based approach suggests that validity is not settled by a single study (e.g., Kane, 2013). Instead, the validity argument for particular score interpretations and uses is stronger or weaker based on available evidence gathered across multiple studies (e.g., Kane, 2013).

Notably, factorial validity and content validity–i.e., domain relevance and coverage (e.g., Messick, 1989)–are required for *any* score interpretation or use (e.g., AERA et al., 2014; Leach et al., 2020), and the framework proposed by Whitehouse and colleagues (2021) addresses both forms. For locally developed school climate surveys lacking an explicit theoretical underpinning (e.g., the JCPS CSS), the Whitehouse et al. framework follows previous studies (e.g., Ramelow et al., 2015) by assessing coverage based on the dimensions and relationships theorized by an existing school climate model.

The necessity of domain relevance explains the ubiquity of factor analysis in existing school climate validity studies (e.g., Aldridge & McChesney, 2021; Bear et al., 2011, 2015, 2018; Grazia & Molinari, 2022; Ryberg et al., 2020; Shukla et al., 2019; Waasdorp et al., 2020; Whitehouse et al., 2021; You et al., 2014; Zullig et al., 2015). Whitehouse and colleagues followed the standard practice of using expert judgment to assess content validity (e.g., AERA et al., 2014; Kane, 2013). However, the transparency of the process and therefore the strength of validity evidence (e.g., *Standards*, Standard 1.9), could be improved by using a standardized rubric to guide independent assessment (e.g., Difazio et al., 2018; Rubio et al., 2003).

Although reliability and validity are typically treated as separate in the literature, validity is undergirded by score reliability and reliability is, in-turn, based on a known factor structure (e.g., AERA et al., 2014; Henson, 2001; Zullig et al., 2015). Therefore Kane (2013) includes internal-consistency reliability as a form of validity evidence, underscoring the importance of analyzing and reporting reliability coefficients to support validity of score interpretations and uses (e.g., Aldridge & Ala'l, 2013). Unfortunately, researchers thus far have not reached consensus on assessing reliability for ordinal scales with more than two but less than five response options (e.g., Muthén, 2013a, 2020; Raykov & Marcoulides, 2011). This may explain why Whitehouse et al. (2021) did not report reliability coefficients for their 4-point Likert-type scale (c.f., Bear et al., 2011, 2015, 2018).

Besides assessing factorial validity, researchers have sought evidence of concurrent (e.g., Bear et al., 2011, 2015, 2018), convergent (e.g., Aldridge & Ala'l, 2013), discriminant (e.g., Aldridge & McChesney, 2020; Bear Yang, Mantz, et al., 2014),

and predictive validity (e.g., Aldridge & Ala'l, 2013), and MI (e.g., Bear et al., 2011, 2015, 2018; Whitehouse et al., 2021, Yang et al., 2013). In general, scholars recommend more comprehensive reliability and validity (e.g., Lenz et al., 2021; Ramelow et al, 2015; Zabek et al. 2022; Zullig et al., 2015) assessment and reporting for school climate measures, including specific calls for evidence of predictive validity (Bear, Yang, Pell et al., 2014) and MI across grades (Whitehouse et al., 2021).

Measurement Invariance and the Alignment Method

A practical concern for educators and policymakers is making within-school subgroup and/or between-school comparisons in school climate ratings (e.g., Schweig et al., 2019). In such cases, measurement invariance (MI) between the groups of interest must be established to ensure such comparisons are meaningful (e.g., Byrne & van de Vijver, 2010; Immekus, 2021; Shukla et al., 2019; Waasdorp et al., 2019; Whitehouse et al., 2021). Interestingly, several reviews of school climate measures omit any mention whatsoever of MI among included instruments (Clifford et al., 2012; Kohl et al., 2013; Lewno-Dumdie et al., 2019; Olsen et al., 2017; Ramelow et al., 2015; Wang & Degol, 2016). Among school climate validity studies that include invariance testing, the majority have sought to establish MI between two or three groups using the same MGCFA approach as Whitehouse et al. (2021); e.g., between middle and high school students (Waasdorp et al., 2020), Black, Hispanic, and White students (Bear et al., 2011), middle school students in Mexico and the U.S. (Shukla et al., 2019), teachers and administrators (You et al., 2014), and between Chinese and American parents (Yang et al., 2021).

Results are mixed, with many studies failing to obtain scalar invariance, i.e., invariant loadings and intercepts (e.g., Asparouhov & Muthén, 2014), across some (e.g.,

Bear et al., 2011; Bear, Yang, Pell, et al., 2014; Shukla et al., 2019; You et al., 2014) or all (e.g., Whitehouse et al., 2021) groups and/or subscales under consideration (c.f., Konold et al., 2021). Reported evidence of strict invariance is practically nonexistent in school climate validity studies (Saint et al., 2021). This scenario is not surprising, given that establishing scalar invariance via MGCFA is unlikely for multi-factor scales (Marsh et al., 2018) and that the procedure is especially onerous when more than two subgroups are being compared (Flake & McCoach, 2018). Flake and McCoach (2018) further suggest that applying MGCFA to Likert-type polytomous items renders model fit interpretations (i.e., identifying MI) tenuous. If testing invariance across multiple groups is challenging even for experienced researchers, the task is likely out-of-reach for most educators, who may not even understand why establishing MI is important (e.g., Schweig et al., 2019).

Asparouhov and Muthén (2014) note that scalar invariance required by MGCFA is rarely obtained when assessing more than two groups. Marsh et al. (2018) suggest the assumption of scalar invariance inherent in the MGCFA approach is untenable in large studies. They lament the tendency in applied research to either rely on chance modification indices to attain partial invariance or ignore invariance altogether. Indeed, the scale examined by Whitehouse et al. (2021) did not exhibit even acceptable configural invariance via MGCFA across three racial/ethnic subgroups (c.f., Bear et al., 2011), yet as discussed above, the authors analyzed group scoring differences anyway. To address the issues inherent in assessing MI through MGCFA, Asparouhov and Muthén (2014; p. 495) developed the alignment method "to estimate group-specific factor means and variances without requiring exact measurement invariance." Flake and

McCoach (2018) applied the alignment method to polytomous items in a simulation study and found that the approach to be helpful for examining MI across many groups when specific conditions were met. Another advantage is that the alignment method permits researchers to compare up to 100 groups (Asparouhov & Muthén, 2014), far surpassing the plausible two or three group comparisons via MGCFA.

Like MGCFA, the alignment method has as its starting point the estimation of a configural CFA model with good fit across all groups (Asparouhov & Muthén, 2014). This model freely estimates item loadings and intercepts, with factor means and variances fixed to (0, 1), respectively (e.g., Immekus, 2021). Because these parameters are not identified, latent factor scores are not comparable across groups, however because item intercepts and loadings are not constrained, the model represents the best possible fit (Asparouhov & Muthén, 2014). The second step of the alignment method is alignment optimization, in which the algorithm seeks to minimize noninvariance across groups while estimating group-specific means and variances (Asparouhov & Muthén, 2014). Optimization is achieved by minimizing a total loss function that represents the weighted (by group size) difference between all pairs of scaled loadings and intercepts between all possible grouping pairs (Asparouhov & Muthén, 2014). The weighting scheme implies that larger groups will have greater influence on the total loss factor.

Asparouhov and Muthén (2014) developed two basic approaches to the optimization stage; *fixed* and *free*. The former constrains a selected group mean and variance (0, 1) whereas the latter constrains only the mean to 0 and freely estimates the variance. Crucially, irrespective of whether the *fixed* or *free* approach is taken, the alignment optimization step does not change model fit achieved by the configural model

(Asparouhov & Muthén, 2014). In simulation studies, Asparouhov and Muthén found that for more than two groups with moderate noninvariance among item parameters (10% to 20%), the *free* option was preferred. The *fixed* option performed well (i.e., parameters exhibited small absolute bias) with a small number (< 60) of groups with at least 100 members and fewer than 20% of item parameters were noninvariant. Several studies (e.g., Byrne & Van de Vijver, 2017; Cieciuch et al., 2018; Immekus, 2021; Marsh et al., 2018) reported model identification problematic when using the *free* approach, and therefore recommended the *fixed* approach. In their simulation study of polytomous items, Flake and McCoach (2018) found adequate parameter recovery using the *fixed* option with default robust maximum likelihood (MLR) estimation with 3, 9, or 15 groups. However, they suggest that the method's automated ad hoc noninvariance testing procedure may require more than 15 groups to adequately measure explained variance (i.e., $R^2 \geq .90$).

Asparouhov and Muthén (2014) warn against cases where the assumption of approximate MI is not met. In a model where only a few item parameters are invariant and most item parameters have a similar amount of noninvariance, the automated procedure may select noninvariant items as invariant. To guard against the alignment producing biased parameters when more than 25% of parameters are noninvariant, they recommend correlating aligned factor means with those produced from a Monte Carlo simulation using real data parameters as starting values (Muthén & Asparouhov, 2014). For the alignment optimization procedure to produce trustworthy results, Muthén and Asparouhov (2014) recommend no more than 25% of item parameters be noninvariant (see also Flake & McCoach, 2018; Asparouhov & Muthén, 2014) and correlations ≥ .98 between aligned factor means and Monte Carlo simulated ordered factor means.

Purpose of the Study and Research Questions

The preceding discussion highlights the need to select a specific intended interpretation and use of school climate survey scores to define and delimit the types of validity evidence to be gathered and evaluated (e.g., AERA et al., 2014; Kane, 2013). As such, this study will focus on the district's stated intention to compare CSS scores across Black and White middle school student groups (JCPS 2018b, 2019a). The purpose of the current study is twofold. First, the study seeks to replicate and extend Whitehouse et al.'s (2021) validity framework by (a) using a standardized rubric to assess content validity, (b) using the alignment method to evaluate measurement invariance of the survey across 30 school x race groups, and (c) analyzing the predictive validity of aligned factor mean scores. As such, the extended validity framework endeavors to answer the following broad research questions:

1. Are the instrument's proposed dimensions supported by empirical evidence?
2. Do the instrument's empirically supported dimensions adequately address each of the domains of the preferred school climate definition and taxonomy?
3. Can score differences between racial/ethnic subgroups on the empirically supported dimensions of the instrument be meaningfully interpreted?
4. To what extent do scores predict academic achievement?
5. To what extent do scores predict behavior outcomes?

Second, the study aims to provide useful evidence to the school district regarding the validity of comparing CSS scores across Black and White middle school student groups and using scores to inform continuous improvement of student learning. Thus, the broad research questions posed above require adaptation for the specific application here.

To accomplish the goals of replicating and extending the Whitehouse et al. (2021) framework while also providing useful CSS validity evidence to JCPS, this study seeks to answer the following research questions:

1. Are the proposed dimensions of the JCPS middle school CSS supported by empirical evidence?
2. Do the empirically supported dimensions of the JCPS middle school CSS adequately address each of the five domains of school climate as defined by the NSCC (2020, 2021)?
3. Can score differences between Black and White middle school students on empirically supported CSS dimensions be meaningfully interpreted?
4. To what extent do CSS scores predict NWEA MAP math and reading percentiles?
5. To what extent do CSS scores predict discipline referrals?

Implications

This study synthesizes and extends prior research by proposing a standardized validity testing framework for locally developed school climate instruments. The framework recognizes the widespread use of practitioner-developed measures that lack theory-grounding and/or supporting reliability and validity evidence, as well as the need for culturally responsive school climate measures. The application of the framework to answer the research questions above will provide practical guidance to JCPS regarding its intention to compare CSS scores between and across middle school racial/ethnic subgroups. Regardless of whether the findings support the intended use or indicate further scale development is necessary, the approach outlined here can also be applied to investigate the validity of numerous other suggested uses of the various CSS versions.

Results will also contribute to the ongoing literature on the psychometrics of practitioner-based measures. Finally, because the proposed validity testing framework is broadly applicable, the implications for both practice and research extend beyond the CSS. Indeed, the framework is not limited to school climate but can be adapted to any practitioner-designed instrument intended to measure a latent construct but lacking a clear theoretical basis and relevant supporting reliability/validity evidence.

METHODS

Participants

JCPS is a large, urban district located in Louisville, KY. In the 2018-19 school year, the district served more than 94,000 students in 168 school sites. That year, 92% of JCPS middle school students (grades 6-8) completed the online CSS between January and March. As per usual district procedures, respondents (n = 20,241; 49% female, 51% male) were given time to complete the CSS during regular school hours. Respondents in 2018-19 were 42% White, 37% Black, 11% Hispanic/Latino, and 10% categorized as Other. Last, roughly 67% were eligible for free or reduced-price lunch (FRL).

Inclusion and Exclusion

A total of 13 middle schools serving special student populations were excluded from the analysis, primarily due to small enrollment sizes. Of the remaining 25 middle schools, an additional 10 schools were excluded. Of those 10 schools, three were part of combined schools (two served grades 6-12 and one served grades K-12), two were gender-isolated, two did not serve all three grades, and three had relatively small student populations. This left a sample of n = 15 middle schools to be included in the study. The BW sample (N = 11,385) included only Black (45%) and White (55%) students from the 15 retained middle schools. Roughly 62% of BW sample respondents were FRL-eligible.

All responses to the 2018-19 online student CSS were mandatory, and thus there were no missing data. However, students could select a *Prefer not to respond* option for each item. Following de Leeuw et al. (2016), these non-substantive responses were

treated as missing. The large number of surveys (n = 4,962) in the BW sample with one or more such responses precluded replacement due to non-randomness, thus only surveys with all substantive responses to the final set of CSS items (more details below) were retained for analyses. This resulted in a final analytic (FA) sample of n = 6,423 students (48% female, 52% male). Students in the FA sample were 58% White and 42% Black, with 59% eligible for free or reduced-price lunch. Table 1 provides demographic information for Black and White CSS respondents in (a) all middle schools, (b) the BW sample, and (c) the FA sample.

Table 1

Demographic Information for Black & White 2018-19 CSS Respondents (Grades 6-8)

Category	All Middle	BW Sample	FA Sample
Number of Students[a]	16,077	11,385	6,423
Female	49%	49%	48%
Male	51%	51%	52%
Black	47%	45%	42%
White	53%	55%	58%
Free/Reduced Lunch	65%	62%	59%

Note. CSS = Comprehensive School Survey; a – Black and White students only.

Power Analysis

Although no formal power analysis procedure exists for the alignment method, researchers (e.g., Asparouhov & Muthén, 2014; Flake & McCoach, 2018; Muthén & Asparouhov, 2018) have provided rough sample size guidelines (rules of thumb) based on simulation studies. For example, Asparouhov and Muthén (2014) obtained acceptable MI results from groups as small as n = 100, although they suggest that much larger

sample sizes may be necessary in some (undescribed) scenarios. Asparouhov and Muthén (2018) also suggest that the Alignment Method is best for fewer than 100 groups and is preferable to a multilevel modeling approach (e.g., De Jong et al., 2007), when there are fewer than 30 groups and/or a small number of items per factor. For surveys with polytomous items, Flake and McCoach suggest that 50 or more groups may be required to assess MI using a multilevel modeling approach. Thus, the Alignment Method seems appropriate for the 30 school x race groups (M = 214.10, n_{min} = 124, n_{max} = 403) in the current sample, based on the number and size of groups included. The proposed item composition of CSS factors (more details below) and the number of groups in the middle school CSS sample also suggest the Alignment Method is preferable to multilevel modeling for assessing MI (see Research Question 3 in the Data Analysis section below).

Measures

Comprehensive School Survey

JCPS has administered the CSS annually to students in grades 4-12, parents, and employees since 1996-97. Currently, there are eight (three student, three employee, one parent English and one parent Spanish) versions of the survey, with slight variations in item composition and/or wording, depending on the intended respondents. The CSS is administered online except that parents may opt to complete a paper survey in English or Spanish. In addition to school climate perspectives, the survey also collects demographic information, although the non-anonymous administration of all CSS student versions in 2018-19 permitted linkage to student data already collected by the district.

Figure 1

Content Validity Rubric

School Climate Definition and Taxonomy

In this study, school climate refers to the quality and character of school life (NSCC, 2021), and is comprised of the aggregated, subjective perceptions of a school community along five domains (dimensions in parentheses): *safety* (rules & norms; physical security; social-emotional security), *interpersonal relationships* (respect for diversity; social support-students; social support-adults), *teaching and learning* (support for learning; social & civic learning), *institutional environment* (school connectedness/engagement, social inclusion, physical surroundings), and *social media* (social media). As such, school climate reflects a school's norms, goals, values, practices, and organizational structures. School climate does not include objective measures (e.g., discipline or attendance rates), ratings of student self-efficacy, or ratings of school administrators.

Click here for more details on the NSCC's (2021) definition and taxonomy of school climate

Item:

Instructions: Please choose the response that best describes the degree to which you feel the item above represents school climate as defined above.

This item is not representative	This item needs major revision to be representative	This item needs minor revision to be representative	This item is representative

Instructions: Please choose the response that best describes the clarity of the item above.

This item is not clear	This item needs major revision to be clear	This item needs minor revision to be clear	This item is clear

Instructions: Please choose school climate domain(s) you feel the item above best represents.

	Safety
	Interpersonal Relationships
	Teaching and Learning
	Institutional Environment
	Social Media
	Unable to classify

Instructions: Please provide any addition comments, suggestions, or revisions for the item above.

Note. Rubric adapted from Difazio et al. (2018); "here" link is to an online summary of the NSCC's (2020, 2021) definition and dimensions of school climate.

The 2018-19 CSS middle school student version consisted of 37 items comprising 14 constructs, or subdimensions, of school climate: *bullying* (1 item), *caring environment* (4 items), *curriculum* (3 items), *home resources* (1 item), *personal safety* (3 items), *school administration* (1 item), *school belonging* (3 items), *school engagement* (3 items), *school resources* (3 items), *self-efficacy* (3 items), *site safety* (1 item), *success skills* (5 items), *teaching* (3 items), and *overall satisfaction* (3 items). All items used a 4-point Likert-type response scale (1 = *strongly disagree*; 2 = *disagree*; 3 = *agree*; 4 = *strongly agree*), with higher scores indicating a more positive climate rating for every item except Item 2 ("At my school, I feel bullying/cyberbullying is a problem"). As discussed above, students were given a non-substantive response option (0 = *prefer not to answer*).

For the current study, all 37 items were independently assessed by the author and another researcher for content validity using a rubric (see Figure 1) adapted from Difazio et al. (2018). Specifically, each item was rated for clarity and representativeness based on the NSCC's (2020, 2021) definition and taxonomy of school climate. Any disagreements were resolved via discussion. Based on this review, nine items were eliminated for non-representativeness, (i.e., both raters scored the item either 'not representative' or 'needs major revision), resulting in a final set of 28 CSS items. These items comprised a total of 11 proposed constructs, or subdimensions. See Table 2 below for item details.

Table 2

2018-19 CSS Middle School Student Version CSS Items (n = 37)

Item	CSS Construct	Retain
1. At my school, I feel bullying/cyberbullying is a problem.	Bullying	Yes
2. I feel my teachers really care about me.	Caring Environment	Yes
3. I believe I can talk with my counselor.	Caring Environment	Yes
4. My school provides a caring and supportive environment for students.	Caring Environment	Yes

Item	CSS Construct	Retain
5. There is at least one adult at my school whom I feel I can trust.	Caring Environment	Yes
6. I have developed more appreciation for music and the arts through courses at my school.	Curriculum	Yes
7. I am able to connect what we learn in my classes to what we learn in other subjects.	Curriculum	Yes
8. The activities (work) my teachers give us really makes me think.	Curriculum	Yes
9. I am very satisfied with my school.	Overall satisfaction	Yes
10. I would rather go to this school than any other school.	Overall satisfaction	Yes
11. I feel safe on my way to and from school.	Personal safety	Yes
12. I feel safe outside the building before and after school.	Personal safety	Yes
13. I feel safe at school.	Personal safety	Yes
14. I really like other students in my school.	School belonging	Yes
15. I feel that I belong in my school.	School belonging	Yes
16. I feel like I am part of my school community.	School belonging	Yes
17. I learn interesting and useful things at school.	School engagement	Yes
18. I enjoy going to school.	School engagement	Yes
19. My classes have a fair number of students in them.	School resources	Yes
20. Textbooks and other school materials are of high quality.	School resources	Yes
21. My school is equipped with up-to-date computers and other technology.	School resources	Yes
22. Adults in my school handle safety concerns quickly.	Site safety	Yes
23. My teacher lets me show what I know in different ways (projects, presentations, tests, etc.).	Success Skills	Yes
24. I feel comfortable stating my opinion in class even if it disagrees with the opinions of other students.	Success Skills	Yes
25. I have opportunities to design and create new solutions, products, or processes	Success Skills	Yes
26. My teachers give me challenging work.	Teaching	Yes
27. My teachers ask us to summarize what we have learned in a lesson.	Teaching	Yes
28. My teachers make me think first, before they answer my questions.	Teaching	Yes
29. I have Internet access at home.	Home Resources	No
30. I am very satisfied with JCPS.	Overall satisfaction	No
31. My principal provides effective leadership at my school.	School Admin.	No
32. I think school is fun.	School engagement	No
33. When I make a decision, I think about what might happen afterwards.	Self-Efficacy	No
34. I accept responsibility for my actions when I make a mistake or get in trouble.	Self-Efficacy	No
35. I do what I believe is right, even if my friends make fun of me.	Self-Efficacy	No
36. I set goals and then work to achieve them.	Success Skills	No
37. My classmates and I have opportunities to work together on projects.	Success Skills	No

Note. CSS = *Comprehensive School Survey*

Academic Outcomes

Since the 2017-18 school year, JCPS has administered NWEA Measures of Academic Progress (MAP) reading and math assessments thrice annually (fall, winter, spring) to all elementary and middle school students. Consistent with prior research linking school climate to standardized test scores (e.g., MacNeil et al., 2009), Spring 2019 MAP math and reading test percentiles were used to investigate the relationship between school climate and academic outcomes (i.e., Research Question 5). Test percentiles rank student scores relative to a nationally normed sample (Thum & Kuhfeld, 2020) and permit interpretable aggregation and comparisons across grade levels. Possible test percentile values ranged from 1-99 and the variable was treated as continuous.

Behavior Outcomes

As per Gage et al. (2016), the relationship between school climate and behavior outcomes (i.e., Research Question 6) was investigated using student discipline referrals to operationalize behavior outcomes. The referral process is subjective and likely varies across schools and student groups. For example, Hollands et al. (2022) observed racial disparities in referrals favoring non-Black versus Black JCPS students. However, given the district's mandate to reduce out-of-school suspensions, I considered referrals a more reliable indicator of behavior than suspensions (e.g., Hollands et al., 2022). Like Hollands et al., I treated referrals as count data.

Psychometrics

Although initial CSS development was guided by a committee appointed by district leadership, there is little to no available information regarding that process, including details about item selection or results from any pilot studies (c.f., AERA et al.,

2014). Unlike other publicly available school climate scales (e.g., Bear et al., 2011, 2015, 2018; Bear, Yang, Mantz, et al., 2014; Clifford et al., 2012; Kohl et al. 2013; Ramelow et al., 2015, Ryberg et al., 2020), JCPS does not currently provide reliability or validity evidence for CSS. Following a major redesign of the survey in 2007-2008, internal (Muñoz, 2008; Muñoz & Lewis, 2009) and external (Rudasill & Rakes, 2008) researchers conducted complementary psychometric analyses of the CSS, however those reports are no longer posted to the JCPS website. Their availability is somewhat moot, given that the CSS was substantially redesigned in 2018 "to eliminate redundant and poorly functioning items" (Lewis, 2019, p. 1). The redesign was intended to reduce respondent burden and increase response rates.

Although JCPS nowhere provides documentation of the criteria used to identify and remove those items, the need for eliminating redundancy largely arose from ad hoc additions of new CSS items over time. Notably, changes to item composition by JCPS Cabinet members are anticipated in the district's annual survey administration process (Lewis, 2019). Indeed, following the 2018 item reduction effort, several new items were added to the CSS in 2019 and again in 2020. As with previous changes to the survey, no information regarding the conceptual or methodological justification for the inclusion of the new items has been made available to the public, although the additional items were largely related to the COVID-19 pandemic.

Data Analysis

Because the JCPS has proposed a dimensional structure for the CSS (see Table 2), my approach deviates slightly from the traditional EFA-CFA approach taken by Whitehouse et al. (2021). As in Leach et al. (2020), this study first analyzed an *a priori*

CFA model and then, if necessary, turned to the traditional EFA-CFA approach to ascertain the underlying CSS factor structure in order to address Research Questions 1, 2, 4, and 5, and to facilitate assessing MI via the Alignment Method (Research Question 3). Brown (2006) warns against conducting EFA and CFA on the same sample, thus the approach here required randomly splitting the FA sample at the school level to create two independent subsamples; FA-EFA (n = 7 schools) and FA-CFA (n = 8 schools) for EFA and CFA, respectively (e.g., Leach et al., 2020). I used an online random number generator (www.random.org) to generate a random set of five unique integers without replacement from the range [1, 15], which resulted in the set {1, 2, 3, 5, 8, 9, 15}. I then used the corresponding anonymized school-level identifiers to create the two subsamples (i.e., students in Schools 1, 2, 3, 5, 8, 9, and 15 constituted the FA-EFA subsample and students in Schools 4, 6, 7, 10, 11, 12, 13, and 14 comprised the FA-CFA subsample). Preliminary data assembly, including random subsample generation, was conducted in R using RStudio version 1.4.1717. The analytic steps taken to address each of the six Research Questions are discussed below.

Research Question 1

To investigate the underlying dimensions of the CSS, I first use the FA-CFA subsample to conduct a CFA of the *a priori* model shown in Table 2 using Mplus 8.1 (Muthén & Muthén, 2017). Although Table 2 shows 11 factors comprised of 28 items, four of those factors (*Bullying*, *Overall Satisfaction*, *School Engagement*, and *Site Safety*) consisted of only one or two items. Costello and Osborne (2005) suggest that factors composed of fewer than three items are not interpretable, thus those four factors were excluded from the CFA. Adequate model fit would provide evidence supporting the

seven retained dimensions (22 items) proposed by JCPS. As per Brown's (2006) recommendation for ordinal (i.e., Likert-type) items, mean and variance adjusted weighted least squares (WLSMV) estimation was used. Model fit was assessed using Comparative Fit Index (CFI), root mean squared error of approximation (RMSEA), Standardized Root Mean Square Residual (SRMR). Adequate model fit was based on Hu and Bentler's (1999) criteria; CFI > 0.95, RMSEA < 0.06, SRMR < 0.08.

Assuming inadequate fit of the *a priori* model, I conducted EFA with default Geomin oblique rotation on the FA-EFA random subsample. Per Costello and Osborne (2005), item loadings ≥ 0.32 on a single factor were considered sufficient and only factors comprised of at least three such items were retained. Eigenvalues ≥ 1 were examined for supporting the number of factors in the preferred model (e.g., Kaiser, 1960), although not treated as a strict rule (e.g., Fabrigar et al., 1999). Muthén (2013b) suggests parallel analysis (e.g., Horn, 1965) is not reliable when using WLSMV estimation. Model fit was assessed using the same criteria (i.e., CFI, RMSEA, and SRMR) described above. Nested models demonstrating adequate fit were compared using the DIFFTEST corrected chi-square test (e.g., Leach et al., 2020). To confirm fit of the preferred EFA model, a CFA based on that model was conducted using the random FA-CFA subsample.

Research Question 2[1]

Assuming an acceptably fitting CSS model with interpretable factors, domain coverage of the items comprising the model's retained factors was determined using

[1] Ideally, researchers would assess internal consistency reliability for scale scores based on the preferred CFA model. Unfortunately, methods for computing reliability coefficients for ordinal scales with 3-4 response choices are not currently available (e.g., Muthén, 2013, 2020; Raykov & Marcoulides, 2011).

independent assessments by the author and another researcher using the rubric shown in Figure 1. Raters assessed domain coverage based on alignment between retained CSS items and the NSCC's (2020, 2021) five domains of school climate (safety, interpersonal relationships, teaching and learning, institutional environment, and social media).

Research Question 3

Measurement invariance was assessed using the Alignment Method (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014) on the full FA sample (n = 6,244) using Mplus 8.1 (Muthén & Muthén, 2017). Based on prior studies (e.g., Flake & McCoach, 2018; Immekus, 2021), the *fixed* option was selected and the group with factor mean nearest zero in the configural CFA model chosen as the referent. Following the alignment optimization step, a 100-replication Monte Carlo simulation (n = 250, 500, 1,000, 2,000, and 3,000 groups) was conducted per Muthén & Asparouhov (2014). Criteria for assessing trustworthiness of the alignment (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014) were (a) less than 25% of item parameters found to be noninvariant, and (b) correlations between population and simulated group means ≥ .98.

Research Questions 4 and 5

Assuming trustworthy aligned CSS factor means, I first assessed the predictive validity of CSS scores by examining the strength of correlations between aligned factor means and three outcome variables: NWEA MAP math and reading test percentiles (Research Question 4) and discipline referrals (Research Question 5). I then estimated three separate two-level (students nested in schools) random intercept multilevel regression models (MLMs), one for each outcome. Specifically, for the math and reading MLMs, I regressed student math or reading test percentiles (Level 1) onto aligned group

CSS factor means (Level 2), controlling for student race and sex, and their interaction. MLMs were estimated in R using package *lme4* (Bates et al., 2015).

For Research Question 5, I regressed discipline referrals (Level 1) onto aligned group CSS factor means (Level 2), with students nested in schools, again controlling for race, sex, and race*sex. For multilevel models with overdispersed (i.e., dispersion parameter > 1) count data outcomes, Rowe (2021) recommends using the R package *glmmTMB( )* (Brooks et al., 2017) with negative binomial (NB2) specification instead of *lme4* (see also Bolker, 2022). Assuming overdispersed referrals (e.g., Hollands et al., 2022), I followed Rowe's suggestion.

Per convention (e.g., Mulawa et al., 2018), Level 2 variables in all models were grand mean centered. Statistical significance for regression coefficients was assessed using $p < 0.05$. Because I was interested in the effects of Level 2 variables (i.e., aligned CSS factor means) on Level 1 student outcomes, I computed design effects (DE) for each model, based on the intraclass correlation coefficient (ICC) as follows:

DE = 1 + (M - 1)*ICC, where M = mean school x race group size.

DE > 1.5 was interpreted as support for a multilevel analytic approach per Lai and Kwok (2015). The general random intercepts regression equation for all three MLMs (i.e., Y_{ij} = math test percentile, MAP reading test percentile, or discipline referrals) was as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}race_{ij} + \beta_{2j}sex_{ij} + \beta_{3j}race_{ij}*sex_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}CSS(1)_{1j} +...+ \gamma_{0k}CSS(k)_{kj} + u_{ij}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$\beta_{3j} = \gamma_{30}$, where *CSS(1)...CSS(k)* are aligned factor means for k retained factors.

RESULTS

Descriptive Statistics

Table 3 reports descriptive statistics for the FA sample (n = 6,423; 42% Black, 58% White) and for the Black (n = 2,685) and White (n = 3,738) subsamples. Female percentages were similar across the sample (48%) and Black (49%) and White (47%) subsamples. As seen in Table 3, Black students on average reported higher scores on 12 (43%) of the 28 CSS items retained in this study. The average score differences (i.e., the mean of the differences between mean Black and White subgroup scores) between the two subgroups for items on which Black students reported higher mean scores (n = 12 items; M_{Diff} = 0.07, SD_{Diff} = 0.06) were lower than those for the items on which White students reported higher scores (n = 16 items; M_{Diff} = 0.10, SD_{Diff} = 0.07). In other words, when Black students rated CSS items higher than White students the margin tended to be smaller than when White students rated CSS items higher than Black students.

On average, White students scored roughly 25 percentile points higher in math, and 23 percentile points higher in reading, than Black students on the Spring 2018-19 NWEA MAP assessment. These large MAP score disparities between Black and White students are consistent with prior nationwide NWEA findings (e.g., Kuhfeld et al., 2021). In terms of behavior, just over 38% of students in the FA sample received at least one discipline referral, however 58% percent of Black students in the sample received one or more referrals compared with 24% of White students. Similar disparities in discipline referrals among Black and White JCPS students were reported by Hollands et al. (2022).

Table 3

Descriptive Statistics for Full Analytic (FA) Sample and Black/White Subsamples

Category	FA Sample (*n* = 6,423)	Black Subsample (*n* = 2,685)	White Subsample (*n* = 3,738)
	n (%)	*n* (%)	*n* (%)
Female	3,079 (48%)	1,308 (49%)	1,771 (47%)
Male	3,344 (52%)	1,377 (51%)	1,967 (53%)
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
MAP Math Percentile	45.15 (29.36)	30.45 (25.05)	55.69 (27.64)
MAP Reading Percentile	50.84 (29.50)	37.28 (27.55)	60.54 (26.91)
Discipline Referrals	2.09 (5.03)	3.68 (6.52)	0.94 (3.11)
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
CSS Item 1	2.29 (1.00)	2.20 (1.04)	2.36 (0.96)
CSS Item 2	2.94 (0.84)	2.85 (0.89)	3.01 (0.79)
CSS Item 3	3.03 (0.83)	3.01 (0.86)	3.04 (0.81)
CSS Item 4	2.87 (0.78)	2.86 (0.81)	2.89 (0.77)
CSS Item 5	3.39 (0.73)	3.33 (0.79)	3.43 (0.68)
CSS Item 6	2.89 (0.92)	2.81 (0.91)	2.94 (0.92)
CSS Item 7	3.00 (0.71)	3.02 (0.73)	2.99 (0.69)
CSS Item 8	3.12 (0.71)	3.15 (0.73)	3.10 (0.70)
CSS Item 9	2.77 (0.85)	2.72 (0.87)	2.81 (0.83)
CSS Item 10	2.62 (1.00)	2.45 (1.03)	2.75 (0.96)
CSS Item 11	3.19 (0.73)	3.14 (0.75)	3.22 (0.72)
CSS Item 12	3.19 (0.71)	3.20 (0.71)	3.18 (0.71)
CSS Item 13	3.05 (0.73)	3.01 (0.75)	3.07 (0.72)
CSS Item 14	2.93 (0.76)	2.90 (0.79)	2.95 (0.73)
CSS Item 15	2.94 (0.79)	2.88 (0.83)	2.98 (0.75)
CSS Item 16	2.90 (0.77)	2.88 (0.79)	2.91 (0.75)
CSS Item 17	2.95 (0.69)	2.96 (0.71)	2.94 (0.68)
CSS Item 18	2.43 (0.89)	2.44 (0.92)	2.42 (0.86)
CSS Item 19	3.13 (0.71)	3.07 (0.74)	3.17 (0.68)
CSS Item 20	2.47 (0.93)	2.57 (0.94)	2.40 (0.91)
CSS Item 21	2.91 (0.86)	2.93 (0.86)	2.89 (0.86)
CSS Item 22	3.04 (0.81)	2.98 (0.84)	3.09 (0.78)
CSS Item 23	3.20 (0.71)	3.19 (0.74)	3.21 (0.68)
CSS Item 24	2.97 (0.84)	3.05 (0.83)	2.91 (0.85)
CSS Item 25	2.97 (0.79)	3.00 (0.81)	2.95 (0.77)
CSS Item 26	3.18 (0.73)	3.21 (0.75)	3.16 (0.72)
CSS Item 27	2.88 (0.77)	2.97 (0.78)	2.81 (0.76)
CSS Item 28	3.00 (0.74)	3.04 (0.76)	2.97 (0.72)

Note. CSS = *Comprehensive School Survey*; MAP = *NWEA Measures of Academic Progress;* M = mean; SD = standard deviation. See Table 2 for CSS item details.

Research Question 1

Demographics of the random FA-EFA (*n* = 2,801, 50% Female, 44% Black) and random FA-CFA (*n* = 3,622; 47% Female, 40% Black) subsamples were similar (+/- 2 percentage points) to those of the overall FA sample (see Table 3 for FA sample details).

A Priori CFA Model

To assess empirical support of the proposed dimensions of the CSS middle school student version, I first conducted CFA of the *a priori* model on the FA-CFA subsample (Table 4, Model M1). Model M1 included the seven factors (22 items) shown in Table 2 comprised of three or more items each; *Caring Environment*, *Curriculum*, *Success Skills*, *Personal Safety*, *School Belonging*, *School Resources*, and *Teaching*. Model M1 fit was adequate, however its covariance matrix was not positive definite due to an inadmissible correlation (*r* = 1.03) between *Curriculum* and *Teaching*. Modification indices suggested correlating error terms for Items 8 and 26. The modification seemed conceptually plausible, given similar language between Item 8 ("The activities (work) my teachers give us really makes me think") and Item 26 ("My teachers give me challenging work").

As seen in Table 4, the 7-factor Model M1a with correlated error terms for Items 8 and 26 demonstrated adequate fit (RMSEA = 0.05; CFI = 0.97; SRMR = 0.03). Although Model M1a had acceptable fit, *Curriculum* was highly correlated with *Teaching* (*r* = 0.94), *Success Skills* (*r* = 0.90), and *School Resources* (*r* = 0.89). The strong correlations suggested that these four factors could be combined into a single factor. Conceptually, these four factors appear to represent a school's *Instructional Environment* and combining them would simplify CSS score reporting for JCPS.

Table 4

A Priori CFA Model Fit Comparisons (FA-CFA Sample)

Model	Factors	Items	χ^2 (*df*)	RMSEA [95% CI]	CFI	SRMR
M1	7	22	Covariance matrix not positive definite			
M1a	7	22	1984.30* (187)	0.05* [0.059, 0.054]	0.97	0.03
M2 (preferred)	4	21	2474.15* (183)	0.06 [0.057, 0.061]	0.96	0.04

Note. CFA = confirmatory factor analysis; *df* = degrees of freedom; RMSEA = root mean square error of approximation; CFI = Comparative Fit Index; SRMR = standardized root mean square residual.
*$p < .05$

Therefore, I analyzed Model M2, a 4-factor version of Model M1a that combined *Curriculum*, *Teaching*, *Success Skills*, and *School Resources* into a single factor called *Instructional Environment*[2]. Given JCPS' intent to reduce respondent burden, Model M2 also eliminated Item 26 rather than correlating error terms with Item 8. As reported in Table 4, the 4-factor, 21-item Model M2 had acceptable model-data fit (RMSEA = 0.059; CFI = 0.96; SRMR = 0.04) and factor correlations ranged from $r = 0.63$ to 0.82. Although the fit criteria slightly supported Model M1a versus M2, the models were not nested, precluding the recommended DIFFTEST model comparison test (Muthén & Muthén, 2017). However, the relative dimensional parsimony (i.e., four versus seven factors) of Model M2 and the inclusion of two factors with at least four items each might be preferred for both analytic (i.e., model identification and/or convergence) and practical (i.e., score reporting) reasons. Thus, although either model might have been chosen based on model fit criteria, for the purposes here, I used Model M2 in the analyses that follow.

[2] Thanks to Tamara Lewis of JCPS for suggesting this name.

Exploratory Factor Analysis

Given an acceptably fitting *a priori* CFA Model M2, I could have bypassed the EFA-CFA model selection procedure altogether. However, M2 omitted six CSS items prior to analysis and combined four factors into a single factor, thus it seemed possible that EFA might support a CFA model retaining more of the 28 items included here. Specifically, I hoped that the EFA-indicated CFA model would include only factors with at least four items each. Therefore, I compared the fit of EFA models with 1-5 factors, omitting the redundant Item 26. Initial analysis revealed that the strongest loading (i.e., $| \geq 0.32 |$) for Item 1 tended to be negative, despite reverse-coded scores. Given that Item 1 was the only negatively-worded item on the CSS, it is likely that some, or perhaps many, respondents did not adjust their responses accordingly. Thus, I also excluded Item 1 from the EFA, resulting in a total of 26 items included.

Table 5

EFA Model Fit Comparisons (FA-EFA Sample)

Model	Factors	Item	χ^2 (*df*)	RMSEA	CFI	SRMR	DIFFTEST[a]
1-factor	1	26	7385.34* (299)	0.09	0.89	0.06	
2-factor	2	26	4388.31* (274)	0.07	0.93	0.05	2-factor
3-factor	3	26	2617.00* (250)	0.06	0.96	0.03	3-factor
4-factor	4	26	2017.28* (227)	0.05	0.97	0.03	4-factor
5-factor	5	26	1413.35* (205)	0.05	.98	0.02	5-factor

Note. EFA = exploratory factor analysis; *df* = degrees of freedom; RMSEA = root mean square error of approximation; CFI = Comparative Fit Index; SRMR = standardized root mean square residual; DIFFTEST = corrected chi square difference test (Muthén & Muthén, 2017). a – column lists the DIFFTEST preferred *n* factor vs. *n-1* factor models. *$p < .05$

As seen in Table 5, the 5-factor EFA model is preferred in terms of fit criteria and DIFFTEST (5-factor vs. 4-factor; $\chi^2(22) = 498.87$, $p < 0.001$). No items loaded ≥ 0.32

onto factor F5, thus I estimated 4-factor CFA Model E4 indicated by EFA. The 21 items shared between Models M2 and E4 had the same primary loadings (see Table 6) and no additional items loaded to *Personal Safety* or *School Belonging*. Crucially, CFA of the 26-item, 4-factor model (FA-CFA sample) did not produce acceptable fit (RMSEA = 0.64, CFI = 0.946, SRMR = 0.04). Because Model M2 demonstrated acceptable fit without modification, I used it throughout the remainder of the study.

Table 6

CSS Model M2 Factor Loadings, Structure Coefficients, and Residual Variances

CSS Item	*Caring Environment*	*Instructional Environment*	*Personal Safety*	*School Belonging*	*RV*
Item 2	**0.77**	(0.63)	(0.56)	(0.62)	0.41
Item 3	**0.65**	(0.53)	(0.47)	(0.52)	0.58
Item 4	**0.85**	(0.70)	(0.61)	(0.68)	0.29
Item 5	**0.63**	(0.52)	(0.45)	(0.50)	0.61
Item 6	(0.44)	**0.53**	(0.33)	(0.36)	0.72
Item 7	(0.63)	**0.76**	(0.48)	(0.51)	0.42
Item 8	(0.53)	**0.65**	(0.40)	(0.43)	0.58
Item 11	(0.59)	(0.51)	**0.82**	(0.54)	0.33
Item 12	(0.57)	(0.50)	**0.79**	(0.53)	0.37
Item 13	(0.65)	(0.56)	**0.90**	(0.60)	0.19
Item 14	(0.51)	(0.43)	(0.43)	**0.64**	0.59
Item 15	(0.67)	(0.56)	(0.56)	**0.84**	0.30
Item 16	(0.71)	(0.60)	(0.59)	**0.89**	0.21
Item 19	(0.49)	**0.60**	(0.37)	(0.40)	0.64
Item 20	(0.49)	**0.60**	(0.37)	(0.40)	0.64
Item 21	(0.50)	**0.60**	(0.38)	(0.41)	0.64
Item 23	(0.59)	**0.72**	(0.45)	(0.48)	0.49
Item 24	(0.47)	**0.58**	(0.36)	(0.39)	0.67
Item 25	(0.57)	**0.70**	(0.44)	(0.47)	0.51
Item 27	(0.52)	**0.64**	(0.40)	(0.43)	0.59
Item 28	(0.57)	**0.70**	(0.44)	(0.47)	0.51
Ordinal $\alpha^{a,b}$	0.81	0.88	0.86	0.82	

Note. CSS = *Comprehensive School Survey*; Standardized (STDYX) loadings in bold; Structure Coefficients in parentheses; *RV* = residual variance; Model estimated using FA-CFA subsample. a – Chalmers (2018) warns against interpreting ordinal α as a reliability statistic; b – Turner et al. (2017) recommend not reporting confidence intervals.

Research Question 2

Based on CFA Model M2, the second research question asked whether the model's four factors and 21 items adequately address the five domains (safety, interpersonal relationships, teaching and learning, institutional environment, and social media) of school climate, as defined by the NSCC (2020, 2021). To assess domain coverage, I relied on the rubric shown in Figure 1. As mentioned earlier, the author and another experienced researcher independently indicated the corresponding NSCC domain(s) for each CSS item. For the purpose of assessing domain coverage, only those domains indicated by both raters for an item were considered to indicate coverage.

Table 7

CSS Model M2 Items and Factors by NSCC School Climate Domains

NSCC Domain	CSS Model M2 Items	CSS Model M2 Factors
Safety	11, 12, 13	*Personal Safety*
Interpersonal relationships	2, 3, 5 14, 15, 16 24	*Caring Environment* *School Belonging* *Instructional Environment*
Teaching and learning	6, 7, 8, 23, 25, 27, 28	*Instructional Environment*
Institutional environment	4 15, 16 19, 20, 21	*Caring Environment* *School Belonging* *Instructional Environment*
Social Media	None	None

Note. CSS = *Comprehensive School Survey*; NSCC = National School Climate Center; Domains indicated by NSCC (2020, 2021).

As seen in Table 7, both independent raters indicated that the items and factors comprising Model M2 reflected elements of four out of five NSCC (2020, 2021) school climate domains, social media excepted. Among M2's factors *Personal safety* exhibited the best alignment, with its three items mapping 1:1 to the NSCC's safety domain. Per the raters, the items comprising *School Belonging* addressed elements of interpersonal

relationships and institutional environment, and Model M2's combined *Instructional Environment* factor reflected those two NSCC domains plus teaching and learning.

From a dimensional standpoint, the three CSS *Personal Safety* items address elements of the NSCC's (2020, 2021) physical and social-emotional security dimension of safety, but not rules and norms. Items from *School Belonging* and *Instructional Environment* reflect elements of social support-students but do not explicitly reflect respect for diversity or social support-adults. *Caring Environment* items, on the other hand, addresses social support-adults but not the other two interpersonal relationship dimensions. The CSS items comprising *School Belonging* and *Caring Environment* reflect school connectedness/engagement from the institutional environment domain, although representation of social inclusion or physical surroundings dimensions is not explicit. *Instructional Environment* addresses only the physical surroundings dimension of the institutional environment domain, along with both dimensions (support for learning, social and civic learning) from the NSCC's teaching and learning domain.

Research Question 3

To examine whether scoring differences on the four subscales of CSS Model M2 can be meaningfully interpreted, I used the alignment method in Mplus to assess MI (e.g., Asparouhov & Muthén, 2014). The prerequisite for alignment is a well-fitting configural model. Model M2 using the full FA sample (n = 6,423) met the established fit criteria (RMSEA = 0.057; CFI = 0.96; SRMR = 0.04). Mplus allows for alignment of multiple factors concurrently (e.g., Marsh et al., 2018), however the covariance matrix for the four factors comprising Model M2 was not positive definite when estimating all four factors (21 items) simultaneously. Therefore, I followed Muthén's (2017) recommendation for

such cases by aligning each factor separately to assess MI and obtain aligned factor scores for the 30 school x race groups under investigation here.

The respective alignment procedures for *Caring Environment*, *Instructional Environment*, and *Personal Safety* produced no errors or interpretability issues. The covariance matrix of *School Belonging*, however, was found not positive definite due to a negative residual variance (*RV*; i.e., Heywood case) on Item 15 for two subgroups. Although Mplus permits fixing *RV*=0 to overcome this issue, Muthén (2014) suggests doing so is suboptimal and instead recommends changing the model. In this case, however, the 3-item *School Belonging* factor was just identified, thus substantive model alterations were not possible. Therefore, despite the purported advantage of the alignment method versus MGCFA in obtaining parameter estimates without reliance on modification indices (Asparouhov & Muthén, 2014; Marsh et al., 2018), I conducted the *School Belonging* alignment with the residual variance of item 15 fixed to zero. Any resulting validity evidence for *School Belonging*, however, was considered weaker than if no modifications were required.

Table 8

CSS Model M2 Alignment Measurement Invariance Results

Item Parameter	Fit Function Contribution	R^2	Groups with Approx. MI	*M*	*SD*	Minimum Est.	Minimum School	Maximum Est.	Maximum School
CarEnv									
Loading									
Item2	-155.96	0.67	30	1.00	0.07	0.85	7W	1.12	13B
Item3	-170.53	0.21	30	1.00	0.11	0.60	5B	1.14	7W
Item4	-177.66	0.08	30	1.00	0.11	0.79	9W	1.18	3B
Item5	-205.39	0.06	29	1.00	0.23	0.06	1B	1.25	14W
Intercept									
Item2	-189.44	0.63	29	0.30	0.14	-0.05	15B	0.59	8W
Item3	-189.08	0.05	28	0.29	0.15	-0.09	10W	0.75	6B
Item4	-177.33	0.59	30	0.29	0.12	0.06	5W	0.59	3B
Item5	-197.42	0.50	29	0.30	0.17	-0.23	4B	0.58	15W
Sum	-1,462.81		235 (98%)						

Item Parameter	Fit Function Contribution	R^2	Groups with Approx. MI	*M*	*SD*	Minimum Est.	Minimum School	Maximum Est.	Maximum School
InstEnv									
Loading									
Item6	-191.21	0.26	30	1.00	0.15	0.74	15B	1.33	9B
Item7	-172.84	0.43	30	1.00	0.11	0.80	1B	1.23	14W
Item8	-193.74	0.29	30	1.00	0.16	0.62	12B	1.26	14W
Item19	-201.28	0.18	30	1.00	0.17	0.57	5B	1.38	15W
Item20	-183.70	0.00	30	1.00	0.13	0.78	13B	1.24	11B
Item21	-197.99	0.24	30	1.00	0.16	0.72	7W	1.27	11B
Item23	-182.08	0.30	30	1.00	0.13	0.80	15W	1.37	5B
Item24	-184.58	0.26	30	1.00	0.14	0.66	9B	1.25	7W
Item25	-164.31	0.40	30	1.00	0.08	0.80	4W	1.12	4B
Item27	-176.89	0.47	30	0.99	0.11	0.81	3B	1.23	11B
Item28	-156.50	0.50	30	1.00	0.07	0.87	3B	1.16	4B
Intercept									
Item6	-246.05	0.07	26	0.32	0.29	-0.22	8B	1.05	10W
Item7	-163.81	0.84	30	0.32	0.08	0.19	4B	0.50	3B
Item8	-194.75	0.50	30	0.32	0.16	-0.10	2W	0.66	6B
Item19	-257.09	0.07	24	0.31	0.32	-0.53	15B	0.54	3B
Item20	-256.35	0.44	26	0.32	0.31	-0.52	6W	0.77	4B
Item21	-273.85	0.32	22	0.33	0.37	-0.44	9W	1.09	7W
Item23	-186.60	0.60	28	0.32	0.13	0.11	1B	0.62	14W
Item24	-204.42	0.59	30	0.31	0.18	-0.18	6W	0.62	11B
Item25	-182.90	0.76	30	0.32	0.13	0.03	6B	0.50	5W
Item27	-223.21	0.52	26	0.32	0.22	-0.05	10W	0.70	13B
Item28	-165.66	0.74	30	0.32	0.09	0.15	3B	0.52	9W
Sum	-4.359.79		632 (96%)						
PersSaf									
Loading									
Item11	-171.31	0.46	30	1.00	0.10	0.80	10W	1.26	1B
Item12	-172.77	0.29	30	1.00	0.11	0.80	7B	1.39	15B
Item13	-167.19	0.66	30	0.99	0.09	0.73	15B	1.22	8W
Intercept									
Item11	-155.74	0.91	30	0.27	0.06	0.15	5B	0.35	6B
Item12	-166.04	0.80	29	0.26	0.10	-0.12	14W	0.43	9B
Item13	-182.78	0.80	29	0.26	0.13	-0.02	13B	0.45	4W
Sum	-1,015.83		178 (99%)						
SchBel									
Loading									
Item14	-148.23	0.81	30	1.00	0.05	0.77	8B	1.09	3B
Item15	-167.72	0.07	30	1.00	0.10	0.84	12W	1.28	5B
Item16	-181.87	0.39	30	0.99	0.14	0.70	15B	1.35	7W
Intercept									
Item14	-157.83	0.88	30	0.11	0.07	-0.06	6W	0.24	8B
Item15	-153.78	0.82	30	0.11	0.06	-0.01	1B	0.26	6W
Item16	-190.82	0.53	29	0.10	0.16	-0.14	9B	0.61	15B
Sum	-1,000.25		179 (99%)						

Note. CSS = *Comprehensive School Survey*; MI = measurement invariance; *M* = mean; *SD* = standard deviation; Est. = estimate; CarEnv = *Caring Environment*; InstEnv = *Instructional Environment*; PersSaf = *Personal Safety*; SchBel = *School Belonging*; B = Black student subgroup; W = White student subgroup.

Table 8 provides alignment results for each of the four factors retained in CSS Model M2. I used the *fixed* option for each of the four separate alignment procedures, specifying the group with factor mean closest to zero as the referent for each respective factor; *Caring Environment* – School 12W(hite), *Instructional Environment* – School 10B(lack), *Personal Safety* – School 5W, and *School Belonging* – School 4W. As seen in Column 4, just over 69% of item parameters were invariant across all 30 groups. Overall, the percentages of invariant parameters ranged from to 96% (*Instructional Environment*) to 99% (*Personal Safety*, *School Belonging*). In total, only 3% of item parameters (mostly intercepts, with one exception) were noninvariant across the four factors. The percentage of parameter noninvariance was well under the 25% threshold suggested by Asparouhov and Muthén (2014) for ensuring trustworthiness of alignment results.

Based on total (i.e., sum of loading and intercept) item fit function contribution shown in Table 8, Column 2, items 2, 28, 11, and 15 contributed the least amount of noninvariance to *Caring Environment*, *Instructional Environment*, *Personal Safety*, and *School Belonging*, respectively, whereas items 5, 21, 13, and 16 contributed the most noninvariance to those factors. The R^2 value shown in Column 3 is a relative measure of invariance based on across-group parameter variations in the configural model, with values closer to 1.00 indicating higher levels of invariance. R^2 values ranged from 0.00 (item 20) to 0.81 (item 28) for loadings and from 0.05 (item 6) to 0.91 (item 11) for item intercepts. With three exceptions (due to rounding) mean factor loadings across the 30 groups for each factor were 1.00 and mean intercepts ranged from -0.23 (4B) to 0.75 (6B), -0.53 (15B) to 1.09 (7W), -0.12 (14W) to 0.45 (4W), and -0.14 (9B) to 0.61 (15B)

for *Caring Environment*, *Instructional Environment*, *Personal Safety*, and *School Belonging*, respectively (see Table 8, Columns 7-10).

Table 9

CSS Model M2 Aligned Factor Means

Caring Environ.		*Instr. Environ.*		*Personal Safety*		*School Belonging*	
School	Mean	School	Mean	School	Mean	School	Mean
12W	0.00	8B	0.08	5W[a]	0.00	14W[a]	0.22
7W[a]	-0.05	10B	0.00	4B	-0.01	10W[a]	0.22
8W	-0.05	8W	-0.01	7W[a]	-0.03	8B	0.16
5W[a]	-0.05	7W	-0.12	11W[a]	-0.04	3B	0.08
10W	-0.11	12W	-0.13	12W	-0.04	8W	0.07
14W[a]	-0.13	12B	-0.15	4W	-0.05	15W[a]	0.05
8B	-0.13	10W	-0.17	8B	-0.08	3W	0.03
3W	-0.17	2B[a]	-0.22	8W	-0.09	7W	0.01
12B	-0.19	3B	-0.24	15W	-0.09	10B[a]	0.01
11W	-0.23	14B	-0.25	12B	-0.12	11W[a]	0.00
3B	-0.24	1B[a]	-0.26	14B	-0.13	4W	0.00
11B	-0.24	4W	-0.27	10W	-0.14	12W	0.00
4W	-0.25	7B	-0.29	15B	-0.17	5B	-0.05
15W	-0.26	5B	-0.30	14W	-0.21	6B	-0.07
13W	-0.28	4B	-0.36	10B	-0.21	12B	-0.10
9W	-0.31	14W	-0.38	3B	-0.22	5W	-0.10
4B	-0.32	15B	-0.40	3W	-0.25	1B	-0.11
7B[a]	-0.33	5W	-0.41	1B	-0.30	14B[a]	-0.11
10B	-0.34	11B[a]	-0.41	9W	-0.31	9W	-0.14
1B	-0.34	3W	-0.45	1W	-0.32	7B	-0.17
13B	-0.36	15W	-0.48	11B[a]	-0.33	11B[a]	-0.20
1W	-0.36	13B	-0.50	5B[a]	-0.37	2B	-0.20
2B	-0.37	6B	-0.51	7B[a]	-0.38	1W	-0.21
14B[a]	-0.39	9W	-0.55	13W	-0.42	4B	-0.21
15B	-0.40	1W[a]	-0.56	2B	-0.43	13W	-0.24
5B[a]	-0.43	2W[a]	-0.57	13B	-0.44	15B[a]	-0.25
2W	-0.44	9B	-0.60	9B	-0.46	9B	-0.27
6W	-0.45	11W[a]	-0.65	6B[a]	-0.53	13B	-0.31
9B	-0.48	6W	-0.66	2W	-0.61	2W	-0.32
6B	-0.48	13W	-0.71	6W[a]	-1.10	6W	-0.33

Note. CSS = *Comprehensive School Survey*; B = Black student subgroup; W = White student subgroup; a – within-school subgroup mean differences were statistically significant ($p < 0.05$).

Table 9 presents aligned CSS Model M2 group factor means for the 30 school x race groups, with the mean scores sorted in descending order for each of the four factors. The MI evidence just discussed suggests that CSS factor score differences between the Black and White subgroups can be meaningfully interpreted. On average, aligned group factor means for *Instructional Environment* were higher for Black middle school student subgroups (M_{IE} = -0.29, SD_{IE} = 0.18) than those of their White middle school student peers (M_{IE} = -0.41, SD_{IE} = 0.22) in the 15 schools included in this study. The opposite was true for *Caring Environment*, *Personal Safety,* and *School Belonging*, with the mean of means favoring White students (M_{CE} = -0.21, SD_{CE} = 0.15; M_{PS} = -0.25, SD_{PS} = 0.29; M_{SB} = -0.05, SD_{SB} = 0.17) versus Black students (M_{CE} = -0.34, SD_{CE} = 0.10; M_{PS} = -0.28, SD_{PS} = 0.16; M_{SB} = -0.12, SD_{SB} = 0.13). Aligned factor means seen in Table 9 were highly correlated with item summed scores for *Caring Environment* (r = .99), *Personal Safety* (r = .99), *Instructional Environment* (r = .98), and *School Belonging* (r = .99).

As seen in Table 9, nearly three-quarters of the top half of *Caring Environment* (and 60% of *Personal Safety* and *School Belonging*) group factor means were from White middle school student subgroups whereas two-thirds of the top half of *Instructional Environment* group factor means were for Black student subgroups. Mplus provides tests of statistical significance (α = 0.05) for aligned factor mean differences, which allowed me to compare within-school scoring differences between the two subgroups. White middle school students reported statistically significantly higher *Caring Environment* than their Black same-school peers in three schools (5, 7, 14). Black students reported significantly higher *Instructional Environment* scores than White students in three schools (1, 2, 11). White students' aligned *Personal Safety* factor mean scores were

significantly higher than their Black same-school in three schools (5, 7, 11), whereas Black students' aligned *Personal Safety* factor means were higher than their White counterparts in School 6. Finally, statistically significant aligned *School Belonging* factor mean differences favored White students in four schools, namely 10, 11, 14, and 15.

Table 10

Alignment Monte Carlo Simulated Factor Mean Correlations

	Group Size				
	n = 250	n = 500	n = 1,000	n = 2,000	n = 3,000
Factor	r^a	r^a	r^a	r^a	r^a
Caring Environment	.88	.93	.97	**.98**	**.99**
Instructional Environ.	.93	.97	**.98**	**.99**	**.99**
Personal Safety	.96	.975	**.99**	**.99**	**.996**
School Belonging	.92	.95	.97	**.99**	**.98**

Note. CSS = *Comprehensive School Survey*; a – correlation between population factor mean and Monte Carlo (n = 100 repetitions) estimates using real data parameters as starting values. Correlations in bold met Muthén and Asparouhov's (2014) criteria for trustworthiness of alignment results ($r \geq .98$).

I followed Munck et al.'s (2018) approach by conducting Monte Carlo simulations to assess potential bias in alignment estimates despite total item parameter noninvariance well below 25% in this study. Muthén and Asparouhov (2014) recommend simulations when >25% of item parameters are noninvariant but Munck et al. recommend them in all cases. As reported in Table 10, results were mixed. Overall, half of the 20 correlations between population and average simulated factor means, including 60% of correlations for *Caring Environment* and *School Belonging*, were below the .98 threshold recommended by Muthén and Asparouhov (2014) for obtaining unbiased group factor means and variances. None of the correlations for simulated group sizes of 250 or 500, those closest to the actual group sizes in the study (M = 214.10, n_{min} = 124, n_{max} = 403), met the criteria. As such, at least some caution is warranted when interpreting group

mean differences on the four CSS factors in Model M2, especially *Caring Environment* and *School Belonging*. However, the alignment MI evidence on the whole supported, though not conclusively so, interpretability of CSS factor mean differences between school x race groups. Therefore, I proceeded with the analyses.

Table 11

Means, Standard Deviations, and Correlations with 95% CIs for MLM Variables

Variable	*M*	*SD*	1	2	3	4	5	6
1. Math	45.15	29.36						
2. Reading	50.84	29.50	.81* [.80, .82]					
3. Referrals	2.09	5.03	-.30* [-.32,-.27]	-.31* [-.33,-.28]				
4. CarEnv	-0.25	0.14	.28* [.26, .31]	.26* [.24, .29]	-.16* [-.18,-.13]			
5. InstEnv	-0.35	0.21	-.08* [-.11,-.06]	-.08* [-.10,-.05]	-.02 [-.00, .04]	.49* [.47, .51]		
6. PerSaf	-0.24	0.22	.11* [.08, .13]	.10* [.07, .12]	-.07* [-.10,-.05]	.67* [.66, .69]	.45* [.43, .47]	
7. SchBel	-0.06	0.15	.36* [.33, .38]	.31* [.29, .33]	-.16* [-.18,-.13]	.67* [.66, .69]	.51* [.49, .53]	.59* [.58, .61]

Note. *CI* = confidence interval; *MLM* = multilevel model; *M* = mean; *SD* = standard deviation; Math, Reading = NWEA MAP math, reading percentiles; CarEnv = *Caring Environment*; InstEnv = *Instructional Environment*; PersSaf = *Personal Safety*; SchBel = *School Belonging*.
*$p < .05$.

Research Question 4 and 5

To assess the predictive validity of aligned CSS factor means, I computed separate multilevel regression models for NWEA MAP math and reading test percentiles (Research Question 4) and discipline referrals (Research Question 5). Prior to estimating MLMs, I examined relationships among the continuous variables included in the MLMs for preliminary evidence of predictive validity. Table 11 shows the means, standard deviations, and correlations for the continuous variables included in MLMs.

Table 12

MLM Selection for MAP Math & Reading Percentiles, Discipline Referrals

			ANOVA	
Model	AIC	BIC	Versus	χ^2 (*df*)
NWEA MAP Math				
Null (no predictors)	60,017	60,037		
Null + L1	58,790	58,824	Null	1230.40* (2)
Null + L1a	58,786	58,826	Null + L1	7.04* (1)
Null + L1a + L2	58,579	58,647	Null + L1a	214.41* (4)
NWEA MAP Reading				
Null (no predictors)	60,282	60,302		
Null + L1	59,215	59,248	Null	1071.60* (2)
Null + L1a	59,214	59,255	Null + L1	2.37 (1)
Null + L1 + L2	59,040	59,101	Null + L1	182.11* (4)
Discipline Referrals				
Null (no predictors)	20,752	20,773		
Null + L1	19,964	19,998	Null	792.28* (2)
Null + L1a	19,950	19,991	Null + L1	15.49* (1)
Null + L1a + L2	19,886	19,953	Null + L1a	72.28* (4)

Note. MLM = multilevel model; L1 = *Race* + *Sex*; L1a = *Race* + *Sex* + *Race***Sex*; L2 = *Caring Environment* + *Instructional Environment* + *School Belonging* + *Personal Safety* (all grand mean centered; AIC = Akaike information criterion; BIC = Bayesian information criterion; ANOVA = analysis of variance; *df* = degrees of freedom.
**p* < .05

All correlations were statistically significant ($p < .01$), with one plausible exception between discipline referrals and *Instructional Environment* factor means. Math and reading test percentiles were highly positively correlated ($r = .81$) and each was moderately negatively correlated with referrals ($r = -.30, -.31$, respectively). Nearly all correlations of interest were in the expected direction (i.e., CSS factor means positively correlated with math and reading and negatively correlated with referrals), providing baseline evidence supporting the predictive validity of aligned CSS factor means for *Caring Environment*, *Personal Safety*, and *School Belonging*. Unexpectedly, however, *Instructional Environment* was negatively correlated with both math and reading test

percentiles. Although the magnitude of the correlation (i.e., effect size) is small, this surprising finding is practically significant and warrants further discussion below.

Table 13

Multilevel Regression Results

	Math Percentile		Reading Percentile		Discipline Referrals	
Parameter[a]	*β*	*SE*	*β*	*SE*	*β*	*SE*
Fixed Effects						
Race (W)	9.15*	1.63	9.10*	1.52	-1.43*	0.14
Sex (M)	-1.20*	0.96	-5.42*	0.65	0.43*	0.08
Race*Sex	2.95*	1.26			0.48*	0.11
Caring Env.	29.96*	7.26	36.47*	7.49	-0.24	0.62
Instructional Env.	-47.06*	5.31	-44.11*	5.37	1.26*	0.47
Personal Safety	-4.70	3.52	-1.29	3.62	-0.47	0.30
School Belonging	54.02*	4.52	46.55*	4.67	-2.52*	0.39
Random Effects	*SD*		*SD*		*SD*	
Intercept	6.43		5.71		0.48	
Residual	24.63		25.62		4.75	
Dispersion Parameter					10.40	

Note. *β* = regression coefficient; *SE* = standard error; W = White; M = Male; *SD* = standard deviation; a – aligned CSS factor means grand mean centered in all models. *$p < .05$.

Next, I obtained ICCs for multilevel regression models (no predictors) for NWEA MAP math (ICC = 0.11, DE = 24.64) and reading (ICC = 0.08, DE = 17.81) test percentiles, and discipline referrals (ICC = 0.03, DE = 6.54). Based on DE > 1.5 for each baseline model, I proceeded with the model estimation process by first adding Level 1 predictors and then Level 2 predictors. As seen in Table 12, statistically significant ANOVAs favored random intercept models with Level 1 and 2 predictors versus models with only Level 1 predictors and null models (no predictors) for MAP math and reading percentiles and discipline referrals. Table 13 provides details the three preferred MLMs, which included the race by sex interaction term for math and referrals, but not for reading (see Table 12 for model comparison results). Referrals were overdispersed (dispersion

parameter = 10.40), supporting my use of *glmmTMB( )* with negative binomial (NB2) specification (e.g., Rowe, 2022). Visual inspection of fitted vs. residual and QQ plots (e.g., Palmeri, 2016) for the preferred MAP models did not reveal violations of homogeneity of variance or normal distribution of residuals assumptions. I used the R package DHARMa (Hartig, 2022) to test assumptions of the preferred discipline referral MLM. Zero-inflation was not detected (p = .76), and neither the QQ plot nor the fitted versus residuals plot revealed statistically significant deviations.

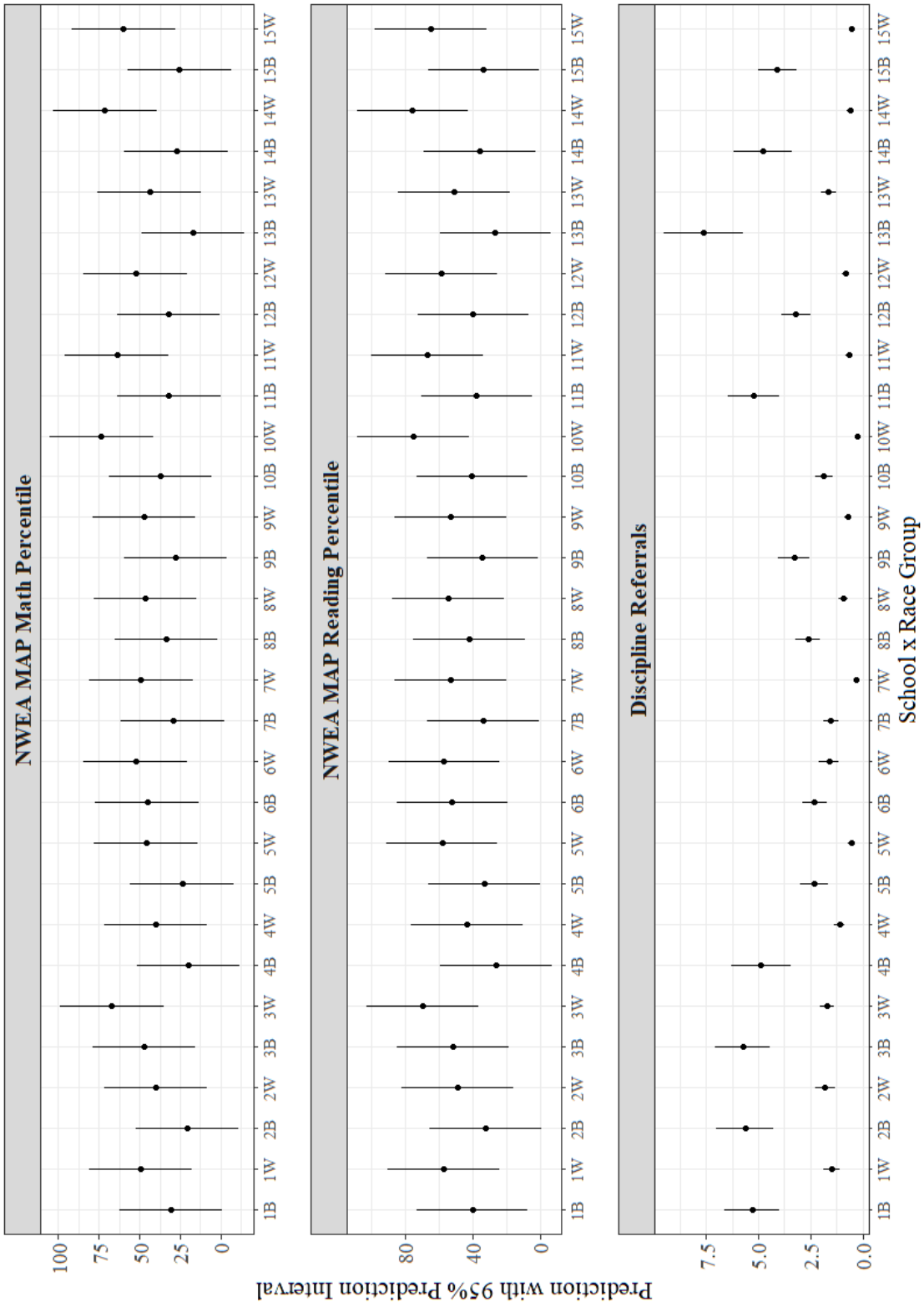
As observed in the pattern of bivariate correlations, statistically significant multilevel regression coefficients in the expected positive direction in the MAP math model supported the predictive validity of *Caring Environment* (β = 29.96, *SE* = 7.26, 95% CI [15.72, 44.22]) and *School Belonging* (β = 54.02, *SE* = 4.52, 95% CI [44.98, 63.04]) aligned group factor means. MAP reading MLM results indicated similar statistically significant predictive validity of aligned group factor means for *Caring Environment* (β = 36.47, *SE* = 7.49, 95% CI [21.76, 51.22]) and *School Belonging* (β = 46.55, *SE* = 4.67, 95% CI [37.16, 55.90]). The negative coefficient in the referrals model indicates that increased *School Belonging* (β = -2.52, *SE* = 0.39, 95% CI [-3.29, -1.75]) was also associated with statistically significantly decreased odds of receiving a discipline referral. The regression coefficients for *Personal Safety* were not statistically distinguishable from zero in any of the three preferred models. Surprisingly, increased perceptions of *Instructional Environment* were statistically significantly associated with lower math (β = -47.06, *SE* = 5.31, 95% CI [-58.01, -36.39]) and reading (β = -44.11, *SE* = 5.37, 95% CI [-55.24, -33.36]) test percentiles, and increased odds of receiving a discipline referral (β = 1.26, *SE* = 0.47, 95% CI [0.34, 2.17]).

Figure 2 shows prediction estimates with 95% prediction intervals for each of the three MLM outcomes (MAP math and reading test percentiles and discipline referrals) for each of the 30 school x race groups. Prediction intervals for MAP math and reading models were obtained via the *predictInterval( )* command in the R package *merTools* (Knowles & Frederick, 2020). Discipline referral prediction intervals were computed manually from standard errors obtained via the *glmmTB( )::predict( )* command.

The regression coefficients reported in Table 13 should not be interpreted in the usual way (i.e., a one unit increase in an aligned group factor mean was associated with a β percentile point change in MAP scores or an Exp(β) change in the odds of receiving a referral). First, the aligned factor means provided by Mplus are not reported in the same units as standard CSS summed or mean factor scores, hindering the clear interpretation of a one unit change (e.g., Munck et al., 2018). Second, the entire range (*maximum value - minimum value*) of aligned group factor means for *Caring Environment* [-0.48, 0.00], *Instructional Environment* [-0.71, 0.08], and *School Belonging* [-0.33, 0.22], was only 0.48, 0.79, and 0.55 units, respectively (see Table 9 for details). As such, the magnitude of the coefficients can be misinterpreted to imply that small changes in CSS group mean scores are associated with large changes in outcomes that are not possible. The purpose here was to assess whether empirical evidence supports the predictive validity of aligned CSS school x race group factor means and not to establish precise relationships with the three outcome variables.

Figure 2

95% Prediction Intervals for MAP Math & Reading Percentiles, and Discipline Referrals



DISCUSSION

Increasing recognition of the importance of school climate (e.g., Bradshaw et al., 2021), combined with federal and state policy requirements to measure the construct (Jordan & Hamilton, 2020) and the fragmented nature of school climate research (Grazia & Molinari, 2022), has resulted in a proliferation of practitioner-developed school climate instruments. Unfortunately, practitioner-developed measures often lack a solid grounding in theory (Ramelow et al., 2015) and/or sufficient reliability and validity evidence to support intended score interpretations and uses (Jordan & Hamilton, 2020; Olsen et al., 2017). Scholars have also emphasized the need to measure school climate equitably across racial/ethnic student groups (Bear et al., 2011; Zabek et al., 2022). To address these gaps, Whitehouse et al. (2021) recently proposed a collaborative (i.e., researchers and practitioners) validity testing framework for practitioner-developed school climate measures with no clear theory-grounding or MI evidence to support the instrument's cultural responsiveness.

Whitehouse et al.'s (2021) suggested approach, however, suffers from the lack of a clearly-defined method for rating item content, reliance on MGCFA to assess MI, and a limited breadth of validity evidence recommended. Therefore, I sought to replicate and extend their validity testing framework for practitioner-developed urban school climate measures by (a) using a standardized rubric to assess content validity, (b) using the alignment method to evaluate MI, and (c) assessing predictive validity. By applying the extended validity testing framework seen in Table 13 to the CSS middle school student

version, I also aimed to provide useful evidence to JCPS regarding the validity of comparing CSS scores across Black and White middle school student groups and using scores to inform continuous improvement of student learning (JCPS, 2018b, 2019a).

Table 14

Validity Testing Framework for Practitioner-Developed Instruments

Steps	Practical Concerns	Methods	*Standards*[a]
1. Settle on a conceptualization that includes a definition, taxonomy, and causal model.	What do we want to measure? What are key aspects? Relationships with other outcomes?	Literature Review	1.1, 1.11, 4.0, 4.1, 4.4, 4.7, 4.8
2. Adapt a content validity rubric based on the definition and taxonomy of school climate from Step 1 and use the rubric to assess item representativeness.	Are our survey questions (items) consistent with the definition in Step 1?	Independent expert ratings	1.9, 1.11, 4.7, 4.8
3. Determine the underlying factor structure of the items retained in Step 2. The recommended is to first test model-data fit of an *a priori* CFA model.	Do our survey questions work as desired? Do we need revisions? Which scores should we report?	CFA, EFA-CFA, ESEM	1.13, 1.14, 1.15, 5.0, 5.1, 5.2
4. Evaluate the internal consistency reliability of all retained factors based on the dimensional structure found in Step 3.	Are our survey scores reliable?	Compute reliability coefficient (e.g., α, ω)	1.14, 2.0, 2.3, 2.19
5. Use the rubric from Step 2 to assess domain coverage of the retained items and factors from Step 3, based on the definition and taxonomy from Step 1.	Are we measuring what we want to measure? All key aspects?	Independent expert ratings	1.9, 1.11, 1.25, 4.12
6. Evaluate measurement invariance (MI) across groups of interest to determine if group mean differences are interpretable.	Should we compare scores between groups?	Alignment, MGCFA, MLM	3.0, 3.1, 3.2, 3.3, 3.6, 3.15, 3.16, 3.17, 4.13
7. Assuming acceptable MI, assess the predictive validity of group factor means on variables of interest included in the causal model from Step 1.	Do our scores matter? How do our scores relate to key outcomes in our model from Step 1?	Correlation, regression, MLM	1.5, 1.16

Note. CFA = confirmatory factor analysis; EFA = exploratory factor analysis; ESEM = exploratory structural equation modeling; MGCFA = multigroup confirmatory factor analysis, MLM = multilevel modeling. a – *Standards* refers to AERA et al. (2014).

Evidence found in this study for sufficient factorial validity and uneven content validity agrees with findings from previous studies of practitioner-developed measures.

Aligned CSS group factor means demonstrated acceptable MI, and both correlational and MLM results supported the expected predictive validity of aligned group mean scores for three of four factors. Overall, the updated framework proposed here (see Table 13) seems advantageous to the original for gathering content, factorial, and predictive validity evidence about practitioner-developed school climate instruments. The mixed results reported above have implications for practice generally, research, and practices at JCPS specifically. Following a discussion of these implications, I highlight limitations of this study and suggest future directions before offering concluding remarks.

Implications for Practice and Research

The mixed findings here, combined with those of myriad other studies of practitioner-developed school climate measures (e.g., Cohen & Thapa, 2017; Hamilton et al., 2019; Ramelow et al., 2015; Zabek et al., 2022; Zullig et al., 2010) suggest that, without expert assistance, practitioners should generally refrain from scale development, including validity testing. Given widely available online tools, creating a school climate survey is easy, but creating one whose scores can be reliably and validly interpreted and used is not, even for experienced scale developers (e.g., Ryberg et al., 2020). When a school climate instrument is not developed through a well-documented, standardized process beginning with a clear conceptualization (e.g., *Standards*, Standard 4.0, 4.1, 4.4, 4.7, 4.8; c.f. Whitehouse et al., 2021), revising the measure in a similarly transparent and regimented approach (e.g., *Standards*, Standard 4.24, 4.25) is not straightforward. In this scenario, assessing the representativeness and domain coverage of the original items is not possible and making ad hoc revisions runs the double risk of removing representative items and adding items that are not necessarily related to the construct the scale is

supposed to measure. Including or adding unrelated items unduly increases respondent burden (e.g., Nathanson et al., 2013) whereas eliminating related items can hinder factorial validity testing. Notably, all of these issues have negatively affected the CSS at one time or another (e.g., Lewis, 2019).

Perhaps the most crucial scale creation mistake inexperienced scale developers (e.g., practitioners) make is beginning without a clear underlying theoretical basis for their construct of interest (c.f., *Standards*, Standard 1.1, 1.11). Some scholars (Kohl et al., 2013; Olsen et al., 2017) recommend that scale developers choose the elements of school climate they wish to measure when creating an instrument. Their suggestion, however, assumes developers have first clearly defined school climate and reviewed existing measures to determine if any are suitable for their intended purpose(s). If domains or dimensions of school climate are unintentionally left unmeasured as a result of unfamiliarity with existing research, rather than intentionally as a result of needs- and research-based decisions, resulting scores may be insufficient for their desired uses (*Standards*, Standard 1.25). Thus, Step 1 in the updated validity testing framework is to settle on a conceptualization of school climate that includes all three elements (definition, taxonomy, and causal model) suggested by Rudasill et al. (2018).

Whitehouse et al. (2021) recommend expert assessment of content validity (i.e., item representativeness and domain coverage), however their validity testing framework provides little practical guidance on the process. The adapted content validity rubric used in Steps 2 and 5 in this study (see Figure 1) is a simple tool that practitioners and scholars alike can use to transparently assess the content validity of school climate instruments (*Standards*, Standard 1.9, 1.11). Notably, the rubric, based on Difazio et al. (2018), is not

limited to school climate measures but can be easily adapted for use with scales intended to measure other latent constructs. Ideally, two or more subject-matter experts will independently rate content validity using the rubric (*Standards*, Standard 4.7, 4.8).

Previous research suggests uneven domain coverage can result when school climate measures lack a solid grounding in theory (Ramelow et al., 2015). Based on independent assessments using the adapted school climate content validity rubric, the 21 items and four factors composing the preferred CSS Model M2 found in Step 3 partially addressed elements of four out of five school climate domains proposed by the NSCC (2020, 2021). Given both the newness and limited theoretical and empirical basis for the NSCC's *Social Media* domain, the failure of the CSS to adequately address the domain is not too concerning. However, Model M2's items and factors did not provide coverage of every dimension of the four more established NSCC domains and three of the model's factors addressed dimensions of multiple NSCC domains. Because the same set of items (plus the redundant Item 26) comprised the acceptably fitting Model M1a, the choice of preferred model did not affect my assessment of CSS item representativeness (Step 2) or domain coverage (Step 5).

I caution here that although prior studies (Ramelow et al., 2015; Whitehouse et al., 2021) used earlier versions of the NSCC's (2020, 2021) conceptualization of school climate as a basis for judging content validity, my choice to use the latest version here in Step 1, though somewhat arbitrary, was necessarily consequential. Despite overlaps between competing school climate definitions (Rudasill et al., 2018), using an alternative conceptualization of school climate as the basis for assessing item representativeness and domain coverage would likely affect the resulting content validity assessment, especially

for measures that were not developed using that particular definition and taxonomy. This underscores the recommendation that all aspects of school climate scale development, including validity studies, should begin with clearly defining the construct (e.g., Chirkina & Khavenson, 2018; Kohl et al., 2013; Lewno-Dumdie et al., 2019; Olsen et al, 2017; Schweig et al., 2019; Whitehouse et al., 2021).

Among the steps in my proposed validity testing framework, settling on a definition and taxonomy of school climate (Step 1) and assessing content validity (Step 2, Step 5) are certainly within the capabilities of most practitioners. However, the more technical aspects such as factor analysis (Step 3, Step 4), examining MI via the alignment method, MGCFA, or MLM (Step 6), and various statistical analyses used to examine predictive validity (Step 7) are unlikely within their purview. Educators' unfamiliarity with, and the technical sophistication of, the psychometric and validity literatures and state-of-the-art methods for assessing reliability and validity (e.g., Schweig et al., 2019; Whitehouse et al., 2021) severely limits their ability to gather and assess reliability and validity evidence. Hence, scholars recommend practitioners partner with experienced researchers to conduct validity studies of practitioner-developed measures (Nathanson et al., 2013; Whitehouse et al., 2021).

Notably, JCPS has in the past followed that recommendation by partnering with university researchers to conduct validity studies of the CSS (Rudasill & Rakes, 2008). Unlike many districts, JCPS employs a team of trained researchers in its Division of Accountability, Research, and Systems Improvement. The Research department has also formally (Muñoz, 2008; Muñoz & Lewis, 2009) and informally investigated the CSS. Unfortunately, however, the district's standard practice of frequently revising the CSS

rendered the findings of those validity studies essentially moot within a few years. If practitioners prefer to regularly modify their school climate instruments, they should (a) follow best scale development practices (e.g., de Leeuw et al., 2014), and (b) conduct validity testing after each update, and (c) document the process and results (e.g., AERA et al., 2014). One danger in the unfortunately common practice of referring to instruments themselves as valid (e.g., Aldridge & Ala'l, 2013; Aldridge & McChesney, 2020; Kohl et al., 2013, Olsen et al., 2017; Whitehouse et al., 2021) is that scale developers (and users) may wrongly assume validity is an intrinsic property of their instruments (c.f., Messick, 1989, 1995; Kane 2013; Schweig et al., 2019) and forgo additional testing when needed, such as after a revision.

Somewhat paradoxically, however, the issues just mentioned do not necessarily imply that locally developed school climate instruments cannot or do not often exhibit at least some good psychometric properties (e.g., Gage et al., 2016; Whitehouse et al., 2021). On the one hand, when scale development is not transparent and the processes of assessing reliability and validity evidence are not standardized and well-documented (e.g., *Standards*, Standard 4.8, 4.10, 4.12, 4.13), appropriate uses and interpretations of scores rely on chance, i.e., on untested or untestable assumptions, rather than evidence (e.g., Jordan & Hamilton, 2020). For example, the 4-point Likert-type response scale used in the CSS, though common (e.g., Immekus, 2021; Whitehouse et al., 2021), precluded assessing internal consistency reliability (e.g., Muthén, 2013a, 2020; Raykov & Marcoulides, 2011) in Step 5, leaving a key factor undergirding validity (AERA et al., 2014; Henson, 2001; Zullig et al., 2015) assumed, but without supporting evidence (c.f. Kane, 2013). Notably, although Gadermann et al. (2012) propose ordinal α as a reliability

coefficient for ordinal items such as those in the CSS, Chalmers (2018) strongly cautions against misinterpreting ordinal α as a measure of observed test score reliability (see also Muthén, 2020).

Along the same lines, four proposed CSS dimensions–*Bullying, School Engagement, Overall Satisfaction,* and *Site Safety*–comprised only one or two school climate-related items each (c.f. Costello & Osborne, 2005), excluding them from further validity testing due to model underidentification. Several other proposed CSS factors comprised exactly three items each. Although those factors contained the minimum of items recommended by Costello and Osborne, model identification can be problematic if an item does not function as expected, as was observed in Step 3 with the *School Belonging* factor in the current study.

On the other hand, the application of a systematic validity testing framework to a practitioner-developed school climate scale may reveal instances where, even when best scale development practices were not followed, at least some assumptions regarding the sound psychometric properties of the instrument were indeed warranted. Prior research has found evidence to support the factorial validity (i.e., underlying simple dimensional structure) of scores from practitioner-developed scales (Gage et al., 2016; Whitehouse et al., 2021). In this study, although the CSS middle school student version lacks a clear theoretical underpinning and well-documented development process (c.f. *Standards,* Standard 4.0), the simple factor structure of both the *a priori* 7-factor Model M1a (22 items) and the modified *a priori* 4-factor Model M2 (21 items) found in Step 3 met the predetermined model fit criteria. The factors in each model exhibited moderate to strong bivariate factor correlations among each other in the expected positive direction.

My proposed Step 4 advocates a traditional approach to assessing factorial validity, namely CFA of an *a priori* model followed by, if necessary, EFA-CFA. The examination of an *a priori* model first recognizes that, as discussed above, practitioner scales can and do exhibit credible evidence of factorial validity (e.g., Whitehouse et al., 2021). From a practical standpoint, educators may have already created score reporting tools aligned with their instrument's proposed factor structure (e.g., JCPS, 2019b), thus evidence supporting that factor structure will eliminate the need to reconfigure any existing tools. From a research perspective, Step 3 (and Step 7) of the validity testing framework also serves to test the theoretical model of school climate selected in Step 1. To the extent that the emergent simple factor structure aligns with the taxonomy (and causal model), the factorial (and predictive) validity evidence supports received theory. Misalignment between the simple factor structure and theorized dimensions may be interpreted to indicate scale revisions are needed.

Practical concerns also underlie my recommendation for a traditional CFA-based approach rather than suggesting exploratory structural equation modeling (ESEM; e.g., Asparouhov & Muthén, 2009). ESEM is often advantageous versus CFA for obtaining a well-fitting model (e.g., Morin et al., 2013), primarily due to ESEM allowing items to load onto multiple factors versus a single factor in CFA (Asparouhov & Muthén, 2009). However, the finding of a well-fitting complex model that is not aligned with a measure's theorized dimensional structure gives little guidance as to the source of misalignment (e.g., Leach et al., 2020). I am unaware of any existing school climate measures whose items are specified to represent multiple factors. Thus, depending on the school climate taxonomy selected in Step 1, the empirical advantage of ESEM may serve to mask poorly

functioning items whose multiple loadings are not consistent with a theorized simple factor structure, hindering assessment of content validity.

ESEM models may also complicate score reporting by precluding subscale scores (e.g., Leach et al., 2020). Although reporting a single, total score is simpler than reporting subscale scores, practitioners may prefer to report subscale scores for the various factors in their measures (e.g., JCPS 2019a, 2019b). This preference agrees with the dimensional focus of prior research investigating the relationships between school climate and student outcomes (e.g., Berkowitz et al., 2017; Daily et al., 2019). In this more nuanced view, a simple structure resulting from CFA or EFA-CFA may be preferable to a complex ESEM factor structure for subscale score reporting.

Of course, CFA is not without shortcomings. For example, reliance on chance modifications to achieve acceptable model-data fit (e.g., correlating error terms as with *School Belonging*) can lead to overfitting a model (Asparouhov & Muthén, 2014; Marsh et al., 2018), weakening the strength of any resulting validity evidence. In general, the analytic approach used to gather factorial validity and reliability evidence (Step 3, Step 4) should align with the complexity of the taxonomy of school climate selected in Step 1, however researchers should also seek to balance empirical and practical concerns. Practitioners, however, may need to indulge researchers' preference to conduct fancy pants analyses (FPAs; Adelson & Owen, 2012) for publication efforts, so that the partnership is mutually beneficial.

To equitably promote a positive school climate, educators need to understand differences in the perceptions of school climate across racial/ethnic student subgroups, necessitating culturally responsive instruments (e.g., Bear et al., 2011; Whitehouse et al.,

2020; Schweig et al., 2019; Zabek et al., 2022). Unfortunately, prior studies using MGCFA (e.g., Whitehouse et al., 2021) and MLM (e.g., Zabek et al., 2022) have failed to establish MI across racial/ethnic subgroups, a critical requirement for meaningfully interpreting group mean differences. Although the Step 6 evidence in this study was not unequivocal, e.g., below-threshold Monte Carlo correlations for smaller group sizes (n = 250, 500), the preponderance supported the trustworthiness of aligned group factor means. Thus, the findings here suggest the alignment method (e.g., Asparouhov & Muthén, 2014) may be a viable alternative to MGCFA or MLM for obtaining unbiased racial/ethnic group factor means without requiring strict MI. However, sample size requirements (more below) may limit the alignment method's applicability to widely administered instruments. Even in large districts like JCPS, non-substantive responses may significantly decrease analytic sample sizes. Practitioners can address this issue by incorporating automated processes aimed at decreasing non-substantive online responses (e.g., de Leeuw et al., 2016).

Multilevel regression coefficients and/or bivariate correlations observed in Step 7 supported the predictive validity of aligned CSS group factor means. For example, I found moderate to strong evidence of the weak to moderate predictive validity of *Caring Environment*, *Personal Safety*, and *School Belonging* in the expected directions for student math and reading test percentiles (e.g., Daily et al., 2019; MacNeil et al., 2009) and discipline referrals (Gage et al., 2016; Huang & Cornell, 2018). Moderate to strong evidence supported weak predictive validity of *Instructional Environment* on those three outcomes, but in the opposite direction than expected based on prior research (e.g., Bear et al., 2011; Gage et al., 2016).

The distinction between the relative strength of predictive validity evidence and the relative strength of predictive relationships supported by that evidence is not trivial. From an argument-based validity perspective (e.g., Kane, 2013), the strength of validity evidence constituting the argument is perhaps more important than the content of the evidence. In other words, strong evidence of weak predictive validity is more convincing than weak evidence for moderate or strong predictive validity, with the caveat that even strong evidence from a single validity study should be supplemented with evidence from additional validity studies (e.g., Kane 2013). The surprising predictive validity findings for *Instructional Environment* also support Kane's assertion that validity arguments are best crafted using evidence gathered from a research program versus a single study.

Perhaps unsurprisingly, given its replicative nature, some results of the current study corroborate findings from the existing literature on the psychometric properties of practitioner-developed instruments. Specifically, previous studies (Gage et al., 2016; Schweig et al., 2019; Whitehouse et al., 2021) have also reported a lack of available reliability reporting, and evidence of uneven content validity (item representativeness and domain coverage. This study adds to the literature by (a) demonstrating that practitioners can create culturally responsive school climate measures that exhibit acceptable levels of MI, and (b) providing evidence supporting the predictive validity of aligned factor means from a culturally responsive practitioner-developed school climate instrument.

The preceding discussion has underscored the position that validity testing of practitioner-developed school climate instruments is best conducted by researcher-practitioner partnerships (Nathanson et al., 2013; Whitehouse et al., 2021). Unfortunately, practitioners may be unaware of the need, and/or unaware of the methods, for validity

testing (Schweig et al., 2019). Thus, Table 13 presents a simple cross-reference tool that connects each step in the updated validity testing framework with practical concerns, plausible methods, and relevant *Standards*. The intent of the tool is to paint a realistic picture of the needs and skill requirements for validity testing to encourage collaborative psychometric investigations of existing practitioner-developed instruments. As such, one potential use of the tool is to facilitate 'real-world' scale development conversations in education leadership courses (e.g., Ed.D., or superintendent or principal certification). I have attempted to make both the framework and tool broad enough to easily adapt for alternate constructs and validity purposes, yet still immediately applicable to the current school climate group mean comparison context. I now turn to the implications for JCPS of applying the framework to the CSS middle school version.

Implications for JCPS

The mixed findings here indicate several areas of improvement for the CSS while also providing validity evidence to support some, but not all, proposed CSS dimensions and JCPS' intention to compare group score differences between Black and White middle school students. As noted above, my somewhat arbitrary Step 1 selection of the NSCC's school climate definition necessarily influenced the validity testing process. However, not all ramifications of the current study for JCPS are definition-dependent. Therefore, I will first delineate which findings were and were not definition-dependent, before drawing out the JCPS-specific implications of the study.

Definition-Dependent

The second step of my proposed validity testing framework involves adapting Difazio et al.'s (2018) content validity rubric based on the school climate definition and

taxonomy selected in Step 1. Thus, any corresponding evidence for (or against) item representativeness and domain coverage is clearly definition-dependent. For example, based on the NSCC's (2020, 2021) school climate definition, I eliminated nine CSS items from the analysis for poor representativeness (Step 2) and later identified uneven domain coverage of the retained items and factors in Model M2 (Step 5). Selecting a different school climate definition and taxonomy as the basis for content validity assessment could result in more (or fewer) items retained and stronger (or weaker) evidence of domain coverage, which in turn could affect factorial validity assessment and scale revisions.

Similarly, the corresponding causal model from Step 1 determines the relevant outcomes for assessing predictive validity in Step 7. However, predictive validity testing is generally less definition-dependent than content validity testing for two reasons; (a) dimensional overlaps among extant school climate models (e.g., Rudasill et al., 2018), and (b) the tendency of researchers to investigate the effects of individual school climate dimensions on various outcomes. To the extent that relationships between *a priori* CSS Model M2 factors and reading/math achievement and referrals are specified in an alternate school climate causal model, the domain to which they belong in that model is irrelevant for assessing predictive but not content validity. That said, the prospect remains that selecting a new school climate model would affect predictive validity results.

Non-Definition-Dependent

Although factorial validity assessment and instrument revision are potentially definition-dependent in terms of item retention, other aspects are likely to be empirically-driven. For example, given the set of representative items retained in Step 2, the evidence supporting the underlying dimensionality and measurement invariance of CSS Model M2

was not definition-dependent. Instead, the validity evidence was affected by decisions to combine highly correlated factors, eliminate redundant items, and correlate error terms due to a Heywood case. Combining the four *Curriculum, Teaching*, *Success Skills*, and *School Resources* factors into *Instructional Environment* for parsimony was also supported conceptually by subject-matter experts at JCPS. The unfortunate inability to assess internal consistency reliability was, as noted earlier, a consequence of using a 4-point Likert-type response option and therefore also not definition-dependent. I now turn to the practical implications of these findings for JCPS.

Next Steps

Evidence from the current study combined with best scale development and validity testing practices from the *Standards* suggests two key CSS improvement tasks; *defining* and *documenting*. Per Step 1 of my proposed validity testing framework, the district should settle on an operational definition of school climate and corresponding taxonomy and underlying causal model. Although I have used the NSCC's (2020, 2021) conceptualization here, there are many options from which JCPS may choose. An advantage of electing to adopt the NSCC's version, however, is that the content validity evidence gathered in this study could be used to guide immediate scale revisions aimed at improving item representativeness and domain coverage. Because the definition, taxonomy, and model underlie all other aspects of scale development and validity testing, the task of defining school climate is paramount[3] (e.g., AERA et al., 2014; Chirkina & Khavenson, 2018; Kohl et al., 2013; Lewno-Dumdie et al., 2019; Olsen et al, 2017;

[3] The CSS appears to measure latent constructs besides school climate. This is not inherently problematic, however, as with school climate, JCPS should clearly conceptualize those constructs and use the extended framework to assess the validity of any intended score uses and interpretations.

Schweig et al., 2019; Whitehouse et al., 2021). Notably, the Step 1 process of defining school climate could entail JCPS discontinuing the CSS and adopting an existing theory-grounded school climate measure (more below).

The second primary task for JCPS is to document all existing (and future) CSS scale development processes. In particular, the process for adding or deleting CSS items should be transparent and aligned with best practices (e.g., AERA et al., 2014; de Leeuw et al., 2014). This task also includes providing any available reliability and validity evidence to support intended CSS score uses and interpretations, which implies also clearly reporting all intended CSS score uses and interpretations (*Standards*, Standard 1.0, 1.1, 1.2). At a bare minimum, all scale development and pertinent reliability and validity evidence should be available to CSS users and respondents (e.g., AERA, 2014). However, given that JCPS is a public school district and all online CSS tools are already publicly available via the district website, I recommend also making all CSS-related scale development and reliable/validity documentation publicly available.

The documenting task also involves providing warnings against potential misinterpretations where various suggested or tool-enabled uses (e.g., comparing group mean scores; JCPS, 2018a, 2018b) are not supported by any evidence (*Standards*, Standard 1.3, 1.4). There are currently no such warnings about any of the many group comparisons enabled via the online CSS Data Tools (JCPS, 2018a), none of which are currently supported by available validity evidence. Crucially, even if JCPS adopts the NSCC's school climate model, alignment MI results from this study are not transferrable or broadly applicable. In other words, evidence supporting the meaningful interpretation of Black and White middle school student group mean CSS score differences *does not*

imply support for any other subgroup comparisons on the CSS middle school version or for any subgroup comparisons whatsoever (including Black and White students) on any other version, or across versions. To reiterate, every single intended or tool-enabled comparison (e.g., JCPS 2018a, 2019a) between two or more groups, including comparing CSS group mean scores across time, grades, or respondent types, must be supported by evidence of MI between those groups. If not, JCPS should provide a warning against possible misinterpretations of those group score differences.

After settling upon a definition, taxonomy, and model of school climate and documenting intended score uses (e.g., group comparisons) and interpretations, JCPS should apply the remaining steps of the updated validity testing framework to obtain relevant content, factorial, and predictive validity evidence for the applicable (more below) intended uses of each version of the CSS. Prior to testing, the district might consider adopting a response option (i.e., two or at least five answer choices) that permits assessing internal consistency reliability. The district may also consider using automated options for decreasing non-substantive responses on its online CSS versions to bolster analytic sample sizes (e.g., de Leeuw et al., 2016). Upon completing the initial round(s) of validity testing, JCPS would find itself facing the ideal adoption, adaption, creation scenario described by Kohl et al. (2013). Specifically, based on the obtained validity evidence, the district could choose to (a) adopt or adapt (i.e., revise) the CSS, (b) adopt or adapt another existing school climate measure that meets its needs, or (c) create a new school climate measure predicated on the definition identified in Step 1.

A word of caution about ‘valid’ or ‘validated’ instruments is appropriate here. Although this process might lead JCPS to adopt a new school climate measure, adopting

a survey with existing validity evidence *does not* eliminate the need for validity testing. Validity evidence for an instrument is context- and sample-specific (e.g., *Standards*, Standard 1.8). The new instrument would require additional supporting evidence based on administration in a new context, namely JCPS (e.g., Kohl et al., 2013). Similarly, scale adaption/revision, and creation also necessitate validity testing (e.g., AERA et al., 2014). In other words, JCPS must conduct additional validity testing irrespective of its decision to adopt, adapt/revise, or create a school climate measure. There are no shortcuts for gathering evidence to support reliable and valid score uses and interpretations. Therefore, given its desire to measure and improve school climate (e.g., JCPS, 2018a, 2018b, 2018c, 2018d; Tatman, 2018, 2019), the district must weigh the available options and determine the most feasible route to achieve that goal. Depending on the capacity of its internal research department, JCPS may need to partner with external researchers to conduct the recommended (e.g., AERA et al., 2014; Kane, 2013) program of CSS validity testing.

Finally, although not a key CSS improvement task, the unexpected negative relationship between *Instructional Environment* and NWEA MAP math and reading percentiles found in Step 7 warrants further investigation. The small, but surprising, effect appears to be related to the tendency for Black JCPS middle school students to report higher perceptions of *Instructional Environment* while also scoring substantially lower on MAP math and reading than their White peers. Crucially, evidence from the alignment of acceptable noninvariance found in Step 6 suggests CSS score differences between the two student groups can be meaningfully interpreted.

The district's recent focus on racial equity offers a potential explanation for the observed effects. In May 2018, recognizing widespread racial disparities in achievement,

opportunities, and discipline outcomes, the Jefferson County Board of Education (JCPS, 2019c) passed the JCPS Racial Education Equity Plan (REP). Among other things, the REP mandates culturally diverse curriculum and instruction, culturally competent professional development, and increased programmatic access for students of color. Elements of each of the four proposed CSS factors (*Curriculum*, *Teaching*, *Success Skills*, and *School Resources*) that were combined into *Instructional Environment* are addressed in the REP. Importantly, many changes specified in the REP had already been implemented prior to the 2018-19 school year investigated in this study (JCPS, 2019c).

Perhaps Black middle school students, on average, perceived *Instructional Environment* more favorably than White students as a result of changes specified in the REP. It is therefore possible that Black students' *Instructional Environment* aligned group factor means served as leading indicators of successful implementation of the REP, foreshadowing later reduced racial/ethnic disparities in math and reading achievement. On the other hand, it is also possible that successful efforts to provide a culturally competent classroom environment were not accompanied by grade-level instruction and/or consistent standards for high quality work. In that case, the negative relationship between *Instructional Environment* and NWEA MAP math and reading percentiles could indicate poor, or at least incomplete, fidelity of REP implementation.

My discussion of *Instructional Environment* focused on NWEA MAP because its predictive validity evidence for achievement was stronger than for behavior. However, the same logic used above also applies to the unexpected positive relationship between *Instructional Environment* and discipline referrals. In other words, the observed positive relationship could indicate successful or incomplete REP implementation. Although the

intervening COVID-19 pandemic response complicates further analysis, findings could have important practical implications for both the CSS and the REP.

Limitations and Future Directions

Although my proposed validity testing framework can be easily adapted for use with a variety of instruments and constructs, it is primarily aimed at examining MI and providing general factorial, content, and predictive validity evidence. The approach seems most suitable for assessing the validity of comparing group mean score differences on measures lacking clear theory-grounding. I applied the framework to a single measure in a single study, however it would appear to be less useful when evidence is needed to support, for example, assigning treatments or classifying schools. Future studies might further extend the framework to incorporate additional types of validity evidence.

The alignment method is a strength for assessing MI, but it may have limited applicability. As noted above, prior school climate validity studies relying on MGCFA (Whitehouse et al., 2021) or MLM (Zabek et al., 2022) failed to establish MI between racial subgroup scores. Aligned CSS group factor means in this study demonstrated acceptable levels of MI, supporting meaningful interpretation of score differences between Black and White JCPS middle school students. The findings suggest alignment may be advantageous for developing culturally responsive school climate measures. Thus, future research should replicate this study by using the alignment method to assess MI between racial subgroup scores on other school climate measures or CSS versions.

As noted earlier, though, below-threshold Monte Carlo simulation results for the two smallest sample sizes in this study (n = 250, 500) suggest the alignment method may not be suitable when group sizes are not adequately large. Crucially, nearly 44% of BW

sample respondents were excluded due to non-substantive responses to one or more CSS items. On the one hand, practitioners should employ methods to reduce such responses (e.g., de Leeuw et al., 2016), thereby increasing analytic sample sizes. On the other hand, uncertainty exists as to whether simulations were necessary in the first place. Asparouhov and Muthén (2014) recommend Monte Carlo simulations to establish the trustworthiness of aligned group factor means when more than 25% of item parameters are noninvariant, which was the case here. However, Munck et al. (2018) suggest simulations are useful at levels of noninvariance below the 25% threshold. In general, more guidance is needed on sample size requirements for unbiased estimation of aligned group factor means and on the amount of noninvariance necessitating Monte Carle simulations to assess parameter bias. Similarly, although researchers have provided general guidelines, or rules of thumb, for choosing between the alignment method and MLM to assess MI (Asparouhov & Muthén, 2018; Flake & McCoach, 2018), future studies should clearly delineate the optimal number and size of groups for using alignment vs. MLM.

Per Schweig et al. (2019), I assessed the predictive validity of aligned CSS factor means using MLM. However, the cross-sectional design severely limits the applicability of results to school improvement efforts. In other words, the findings offer little practical guidance for improving the learning environment without knowing the direction(s) of causal relationships between school climate and achievement and behavior outcomes. Future CSS (and other school climate instrument) validity studies should incorporate experimental designs aimed at determining causal relationships between school climate and student outcomes. Results from such studies would provide stronger evidence to support (or not) using scores to inform school improvement (e.g., JCPS, 2018c).

My proposed validity testing framework advocates a traditional variable-centered approach to examine theorized relationships between school climate dimensions and student academic and behavior outcomes. As such, it follows standard practice (e.g., Howard & Hoffman, 2017) by using factor analytic (CFA, EFA, ESEM), correlation, and regression techniques to obtain predictive validity evidence. The variable-centered approach is well-suited to the NSCC's (2020, 2021) model of school climate and the specific research questions asked in this study.

However, a variable-centered approach may not always be preferred. For example, the theoretical approach and corresponding methodology used in the framework may require adaptation if the systems view of school climate (SVSC; Rudasill et al., 2018) is selected in Step 1. Because the SVSC focuses on patterns of proximal and distal interactions (Rudasill et al., 2018), a person-centered approach may be more appropriate for validity testing. In a person-centered approach, 'persons' or units (e.g., schools) are viewed as systems that can be grouped by similar response patterns. Latent profile analysis (LPA; e.g., Nylund et al., 2007) is a common method used to identify those patterns, or profiles (e.g., Howard & Hoffman, 2017). LPA could be used to identify emergent patterns of school climate and investigate relationships between latent profiles and SVSC-theorized distal outcomes. Additionally, latent transition analysis could be used to observe longitudinal transitions from one latent school climate profile to another (e.g., Leach et al., 2021). A better understanding of the common shifts in patterns of school climate could greatly inform school improvement efforts. Future validity studies based on the SVSC could also identify necessary adaptations of the current framework or propose alternate validity testing frameworks for use with person-centered approaches.

Conclusion

Researchers and practitioners agree on the importance of equitably measuring and improving school climate. However, a research-practice gap in instrument development and validity testing persists between the two groups. Whitehouse et al. (2021) sought to bridge the research-practice gap by proposing a collaborative validity testing framework for practitioner-developed school climate measures lacking theory-grounding. Their framework, however, suffers from key limitations regarding the transparency of content validity assessment, breadth of validity evidence reported, and methods used to examine MI. Therefore, this study sought to replicate and extended their framework by using a standardized rubric to assess content validity, examining MI via the alignment method, and analyzing the predictive validity of aligned group factor scores. I also attempted to further bridge the research-practice gap by cross-referencing each step in the extended framework with practical concerns, relevant methods, and best practices (i.e., *Standards*).

Results suggest the extended framework is superior to the original for assessing MI across racial/ethnic subgroups and obtaining baseline content, factorial, and predictive validity evidence, provided the number and size of groups are adequate. Findings indicate the CSS middle school student version is culturally responsive, although the survey may not sufficiently address all key school climate dimensions. To improve the survey, JCPS must settle on a clear definition and taxonomy of school climate to facilitate a program of validity testing, and publicly document all validity evidence. Future studies should clarify alignment sample size and simulation study requirements and extend the framework to assess additional validity concerns and for use with person-centered approaches.

REFERENCES

Adelson, J. L., & Owen, J. (2012). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy*, *49*(2), 152-162. https://doi.org/10.1037/a0023990

Aldridge, J., & Ala'I, K. (2013). Assessing students' views of school climate: Developing and validating the What's Happening In This School? (WHITS) questionnaire. *Improving Schools, 16*(1), 47–66. https://doi.org/10.1177/1365480212473680

Aldridge, J. M., & McChesney, K. (2021). Parents' and caregivers' perceptions of the school climate: Development and validation of the Parent and Caregiver Survey (PaCS). *Learning Environments Research, 24*(1), 23-41. https://doi.org/10.1007/s10984-020-09308-z

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Asparouhov, T. & Muthén, B. (2009) Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Asparouhov. T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. https://doi.org/10.1080/10705511.2014.919210

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bear, G. G., Gaskins, C., Blank, J., & Chen, F. F. (2011). Delaware School Climate Survey-Student: Its factor structure, concurrent validity, and reliability. *Journal of School Psychology, 49*(2), 157-174. https://doi.org/10.1016/j.jsp.2011.01.001

Bear, G. G., Yang, C., Chen, D., He, X., Xie, J. S., & Huang, X. (2018). Differences in school climate and student engagement in China and the United States. *School Psychology Quarterly, 33*(2), 323-335. http://dx.doi.org/10.1037/spq0000247

Bear, G., Yang, C., Mantz, L., Pasipanodya, E., Hearn, S., & Boyer, D. (2014). *Technical manual for Delaware School Survey: Scales of school climate, bullying victimization, student engagement, and positive, punitive, and social emotional learning techniques*. Newark, DE: Funded by the Delaware Positive Behavior Support Project at the Center for Disability Studies at University of Delaware and Delaware Department of Education. https://wh1.oet.udel.edu/pbs/wp-content/uploads/2011/12/Delaware-School-Survey-Technical-Manual-Fall-2016.pdf

Bear, G. G., Yang, C., & Pasipanodya, E. (2015). Assessing school climate: Validation of a brief measure of the perceptions of parents. *Journal of Psychoeducational Assessment, 33*(2), 115-129. https://doi.org/10.1177/0734282914545748

Bear, G. G., Yang, C., Pell, M., & Gaskins, C. (2014). Validation of a brief measure of teachers' perceptions of school climate: Relations to student achievement and suspensions. *Learning Environments Research, 17*(3), 339-354. https://doi.org/10.1007/s10984-014-9162-1

Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research, 87*(2), 425-469. https://doi.org/10.3102/0034654316669821

Bolker, B., et al. (2022, October 5). *GLMM FAQ*. http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#zero-inflation

Bradshaw, C. P., Cohen, J., Espelage, D. L., & Nation, M. (2021). Addressing school safety through comprehensive school climate approaches. *School Psychology Review, 50*(2-3), 221-236. https://doi.org/10.1080/2372966X.2021.1926321

Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler, and Benjamin M. Bolker. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*(2), 378-400. https://doi.org/10.32614/RJ-2017-066.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford publications.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107-132. https://doi.org/10.1080/15305051003637306

Catalano, R. F., Haggerty, K. P., Oesterle, S., Fleming, C. B., & Hawkins, J. D. (2004). The importance of bonding to school for healthy development: Findings from the Social Development Research Group. *Journal of School Health, 74*(7), 252-261. https://doi.org/10.1111/j.1746-1561.2004.tb08281.x

Chalmers R. P. (2018). On misconceptions and the limited usefulness of Ordinal Alpha. *Educational and Psychological Measurement, 78*(6), 1056–1071. https://doi.org/10.1177/0013164417727036

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language.* Routledge.

Chirkina, T. A., & Khavenson, T. E. (2018). School climate: A history of the concept and approaches to defining and measuring it on PISA questionnaires. *Russian Education & Society, 60*(2), 133–160. https://doi.org/10.1080/10609393.2018.1451189

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research, 47*(4), 665-686. https://doi.org/10.1177/0049124117701478

Clifford, M., Menon, R., Gangi, T., Condon, C., & Hornung, K. (2012). *Measuring school climate for gauging principal performance: A review of the validity and reliability of publicly accessible measures.* American Institutes for Research. https://doi.org/10.1037/e572172012-001

Cohen, J. (2013). Creating a positive school climate: A foundation for resilience. In S. Goldstein, R. Brooks (Eds.), *Handbook of resilience in children* (2nd ed., pp. 411-423). Springer. https://doi.org/10.1007/978-1-4614-3661-4_24

Cohen, J. (2017). School climate, social emotional learning, and other prosocial "camps": Similarities and a difference. *Teachers College Record*. https://www.tcrecord.org/Content.asp?ContentId=22165

Cohen, J., McCabe, L., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *Teachers College Record, 111,* 180-213.

Cohen, J., & Thapa, A. (2017). School climate improvement: What do U.S. educators believe, need and want? *International Journal on School Climate and Violence Prevention, 2*(1), 90-116. https://www.ijvs.org/files/IJSCVP-3-July-2017/4-Cohen-Thapa.pdf

Cornell, D., Huang, F., Konold, T., Shukla, K., Malone, M., Datta, P., Jia, Y., Stohlman, S., Burnette, A. and Meyer, J. P. (2017). *Development of a standard model for school climate and safety assessment: Final report*. Charlottesville, VA: Curry School of Education, University of Virginia. https://www.ojp.gov/pdffiles1/ojjdp/grants/251102.pdf

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), 7. https://doi.org/10.7275/jyj1-4868

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Erlbaum.

Daily, S. M., Mann, M. J., Kristjansson, A. L., Smith, M. L., Zullig, K. J. (2019) School climate and academic achievement in middle and high school students. *Journal of School Health, 89*, 173-180. DOI: 10.1111/josh.12726

De Jong, M. G., Steenkamp, J. B. E., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*(2), 260-278. https://doi.org/10.1086/518532

de Leeuw, E. D., Hox, J. J., & Boevé, A. (2016). Handling do-not-know answers: Exploring new approaches in online and mixed-mode surveys. *Social Science Computer Review, 34*(1), 116–132. https://doi.org/10.1177/0894439315573744

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (Eds.). (2014). *International handbook of survey methodology*. European Association of Methodology.

Difazio, R. L., Strout, T. D., Vessey, J. A., & Lulloff, A. (2018). Item generation and content validity of the Child-Adolescent Bullying Scale. *Nursing Research, 67*(4), 294–304. https://doi.org/10.1097/NNR.0000000000000283

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Finch, H., French, B. F., & Immekus, J. C. (2016). *Applied psychometrics using SPSS and AMOS*. Information Age Publishing.

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 56-70. https://doi.org/10.1080/10705511.2017.1374187

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation, 17*(1), 3. https://doi.org/10.7275/n560-j767

Gage, N. A., Larson, A., Sugai, G., & Chafouleas, S. M. (2016). Student perceptions of school climate as predictors of office discipline referrals. *American Educational Research Journal, 53*(3), 492-515. https://doi.org/10.3102/0002831216637349

Grazia, V., & Molinari, L. (2022). The Multidimensional School Climate Questionnaire (MSCQ) parent-version: Factorial structure and measurement invariance. *International Journal of School & Educational Psychology, 10*(2), 243-247. htpps://doi.org/10.1080/21683603.2020.1828205

Hamilton, L., Doss, C., & Steiner, E. (2019). *Teacher and principal perspectives on social and emotional learning in America's schools: Findings from the American Educator Panels.* RAND Corporation. https://doi.org/10.7249/RR2991

Hartig, F. (2022). DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models. R package version 0.4.6. http://florianhartig.github.io/DHARMa/

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189. https://doi.org/10.1080/07481756.2002.12069034

Hollands, F. M., Leach, S. M., Shand, R., Head, L., Wang, Y., Dossett, D., Chang, F., Yan, B., Martin, M., Pan, Y., & Hensel, S. (2022). Restorative Practices: Using local evidence on costs and student outcomes to inform school district decisions about behavioral interventions. *Journal of School Psychology, 92*, 188–208. https://doi.org/10.1016/j.jsp.2022.03.007

Hopson, L., & Lee, E. (2011). Mitigating the effect of family poverty on academic and behavioral outcomes: The role of school climate in middle and high school. *Children and Youth Services Review*, *33*(11), 2221-2229. https://doi.org/10.1016/j.childyouth.2011.07.006

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185. https://doi.org/10.1007/BF02289447

Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement*. Policy Analysis for California Education, PACE. https://files.eric.ed.gov/fulltext/ED574847.pdf

Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods, 21*(4), 846-876. https://doi.org/10.1177/1094428117744021

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Huang, F. L., & Cornell, D. G. (2016). Multilevel factor structure, concurrent validity, and test-retest reliability of the high school teacher version of the Authoritative School Climate Survey. *Journal of Psychoeducational Assessment, 34*(6), 536-549. https://doi.org/10.1177/0734282915621439

Huang, F. L., & Cornell, D. (2018). The relationship of school climate with out-of-school suspensions. *Children and Youth Services Review, 94*, 378–389. https://doi.org/10.1016/j.childyouth.2018.08.013

Huang, F. L., Cornell, D. G., Konold, T., Meyer, J. P., Lacey, A., Nekvasil, E. K., Heilbrun, A., & Shukla, K. D. (2015). Multilevel factor structure and concurrent validity of the teacher version of the Authoritative School Climate Survey. *Journal of School Health, 85*(12), 843-851. https://doi.org/10.1111/josh.12340

Immekus, J. C. (2021). Multigroup CFA and alignment approaches for testing measurement invariance and factor score estimation: illustration with the Schoolwork-Related Anxiety Survey across countries and gender. *Methodology, 17*(1), 22-38. https://doi.org/10.5964/meth.2281

Jefferson County Public Schools. (2018a). *Comprehensive School Survey.* https://www.jefferson.kyschools.us/departments/data-management-research/comprehensive

Jefferson County Public Schools. (2018b). *What can CSS results tell you?* https://www.jefferson.kyschools.us/file/17901

Jefferson County Public Schools. (2018c). *The pillars of JCPS.* https://www.jefferson.kyschools.us/pillars-jcps

Jefferson County Public Schools. (2018d). *School climate and culture.* https://www.jefferson.kyschools.us/department/school-climate-and-culture

Jefferson County Public Schools. (2019a). *CSS district comparison report 2019.* https://www.jefferson.kyschools.us/departments/data-management-research/comprehensive

Jefferson County Public Schools. (2019b). *Jefferson County Public Schools 2018-2019 Comprehensive School Survey (CSS) results.* https://www.jefferson.kyschools.us/sites/default/files/css2019_all_M.pdf

Jefferson County Public Schools. (2019c). *Jefferson County Public Schools Racial Educational Equity Plan 2018-2020.* https://www.jefferson.kyschools.us/sites/default/files/DistrictRacialEquityPlan%201-8-19.pdf

Jordan, P. W., & Hamilton, L. S. (2020). *Walking a fine line: School climate surveys in state ESSA plans*. FutureEd. Georgetown University. https://www. future-ed. org/wp-content/uploads/2020/01/FutureEdSchoolClimateReport.pdf

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151. https://doi.org/10.1177/001316446002000116

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). American Council on Education and Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42,* 448-457. DOI: 10.1080/02796015.2013.12087465

Knowles, J., & Frederick, C. (2020, June 22). *Prediction intervals from merMod objects.* https://cran.r-project.org/web/packages/merTools/vignettes/Using_predictInterval.html

Kohl, D., Recchia, S., & Steffgen, G. (2013). Measuring school climate: An overview of measurement scales. *Educational Research, 55,* 411-426. https://doi.org/10.1080/00131881.2013.844944

Konold, T. R., & Cornell, D. (2015). Measurement and structural relations of an authoritative school climate model: A multi-level latent variable investigation. *Journal of School Psychology*, *53*(6), 447–461. https://doi.org/10.1016/j.jsp.2015.09.001

Konold, T. R., Edwards, K. D., & Cornell, D. G. (2021). Longitudinal measurement invariance of the Authoritative School Climate Survey. *Journal of Psychoeducational Assessment, 39*(6), 651-664. https://doi.org/10.1177/07342829211011332

Kuhfeld, M., Condron, D. J., & Downey, D. B. (2021). When does inequality grow? A seasonal analysis of racial/ethnic disparities in learning from kindergarten through eighth grade. *Educational Researcher, 50*(4), 225–238. https://doi.org/10.3102/0013189X20977854

Lai, M. H., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The “design effect smaller than two” rule. *Journal of Experimental Education, 83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Leach, S. M., Immekus, J. C., French, B. F., & Hand, B. (2020). The factorial validity of the Cornell Critical Thinking Tests: A multi-analytic approach. *Thinking Skills and Creativity, 37*, 100676. https://doi.org/10.1016/j.tsc.2020.100676

Leach, S.M., Mitchell, A.M., Salmon, P., & Sephton, S.E. (2021). Mindfulness, self-reported health, and cortisol: A latent profile analysis. Journal of Health Psychology, 26(14), 2719-2729. https://doi.org/10.1177/1359105320931184

Lee, T., Cornell, D., Gregory, A., & Fan, X. (2011). High suspension schools and dropout rates for black and white students. *Education and Treatment of Children, 34*(2), 167-192. https://www.jstor.org/stable/42900581

Lenz, A. S., Rocha, L., & Aras, Y. (2021). Measuring school climate: A systematic review of initial development and validation studies. *International Journal for the Advancement of Counselling, 43*(1), 48-62. https://doi.org/10.1007/s10447-020-09415-9

Lewis, T. (2019). *Process analysis narrative: Comprehensive School Surveys 2018-2019.* Jefferson County Public Schools.

Lewno-Dumdie, B., Mason, B., Hajovsky, D., & Villeneuve, E. (2020). Student-report measures of school climate: A dimensional review. *School Mental Health, 12*, 1-21. https://doi.org/10.1007/s12310-019-09340-2

Lindstrom Johnson, S., Reichenberg, R. E., Shukla, K., Waasdorp, T. E. & Bradshaw, C. P. (2019). Improving the measurement of school climate using item response theory. *Educational Measurement: Issues and Practice, 38*(4), 99-107. https://doi.org/10.1111/emip.12296

Long, C. (2017, February 1). *Keeping schools safe, happy places for all students.* National Education Association. https://www.nea.org/advocating-for-change/new-from-nea/keeping-schools-safe-happy-places-all-students

MacNeil, A. J., Prater, D. L., & Busch, S. (2009). The effects of school culture and climate on student achievement. *International Journal of Leadership in Education, 12*(1), 73-84. https://doi.org/10.1080/13603120701576241

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524-545. http://dx.doi.org/10.1037/met0000113

Marx, S., & Byrnes, D. (2012). Multicultural school climate inventory. *Current Issues in Education, 15*(3). http://cie.asu.edu/ojs/index.php/cieatasu/article/view/960

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13-103). Macmillan Publishing Co, Inc; American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741. https://doi.org/10.1037/0003-066X.50.9.741

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory Structural Equation Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 395-436). Information Age Publishing.

Mulawa, M. I., Reyes, H. L. M., Foshee, V. A., Halpern, C. T., Martin, S. L., Kajula, L. J., & Maman, S. (2018). Associations between peer network gender norms and the perpetration of intimate partner violence among urban Tanzanian men: A multilevel analysis. *Prevention Science, 19*(4), 427–436. https://doi.org/10.1007/s11121-017-0835-8

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research, 47*(4), 687-728. https://doi.org/10.1177/0049124117729691

Muñoz, M. (2008). *Exploration and assessment of the reliability of the Comprehensive School Surveys*. Jefferson County Public Schools.

Muñoz, M., & Lewis, T. (2009). *Comprehensive School Surveys (2008-09): Strengthening organizational culture.* Jefferson County Public Schools.

Muthén, B. O. (2017, November 23). *Alignment method question [9:50am]*. Posted to http://www.statmodel.com/discussion/messages/9/24842.html?1520389172

Muthén, B. O. (2020, March 24). *Reliability measures [11:18am]*. Posted to http://www.statmodel.com/discussion/messages/23/625.html?1585187989

Muthén, B. O., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*(978), 1-7. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, B. O., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups. *Sociological Methods & Research, 47*(4), 637-664. https://doi.org/10.1177/0049124117701488

Muthén, L. K. (2013b, March 6). *Parallel analysis for categorical data [9:50am]*. Posted to http://www.statmodel.com/discussion/messages/8/11966.html?1504133952

Muthén, L. K. (2013a, April 16). *Reliability measures [1:09pm]*. Posted to http://www.statmodel.com/discussion/messages/23/625.html?1585187989

Muthén, L. K. (2014, June 27). *Negative Residual Variance [11:05am]*. Posted to http://www.statmodel.com/discussion/messages/11/555.html?1358188287

Muthén, L.K. and Muthén, B.O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Nathanson, L., McCormick, M., Kemple, J. J., & Sypek, L. (2013). *Strengthening assessments of school climate: Lessons from the NYC School Survey*. Research Alliance for New York City Schools. https://files.eric.ed.gov/fulltext/ED543180.pdf

National School Climate Center (2020). *The 14 dimensions of school climate measured by the CSCI.* https://schoolclimate.org/wp-content/uploads/2021/05/NSCC_14-CSCI.pdf

National School Climate Center. (2021). *What is school climate and why is it important?* https://www.schoolclimate.org/school-climate/

National School Climate Council. (2007). *The School Climate Challenge: Narrowing the gap between school climate research and school climate policy, practice guidelines and teacher education policy.* https://schoolclimate.org/wp-content/uploads/2021/05/school-climate-challenge-web.pdf

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling 14*(4): 535–569. https://doi.org/10.1080/10705510701575396

Olsen, J., Preston, A., Algozzine, B., Algozzine, K., & Cusumano, D. (2017). A review and analysis of selected school climate measures. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 91*, 1-12. https://doi.org/10.1080/00098655.2017.1385999

Palmeri, M. (2016). *A language, not a letter: Learning statistics in R. Chapter 18: Testing the assumptions of multilevel models*. https://ademos.people.uic.edu/Chapter18.html

Ramelow, D., Currie, D., & Felder-Puig, R. (2015). The assessment of school climate: Review and appraisal of published student-report measures. *Journal of Psychoeducational Assessment, 33*(8), 731-743. https://doi.org/10.1177/0734282915584852

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Reaves, S., McMahon, S. D., Duffy, S. N., & Ruiz, L. (2018). The test of time: A meta-analytic review of the relation between school climate and problem behavior. *Aggression and Violent Behavior, 39*, 100-108. https://doi.org/10.1016/j.avb.2018.01.006

Rebelez, J. L., & Furlong, M. J. (2013). Psychometric support for an abbreviated version of the California School Climate and Safety Survey. *International Journal of School & Educational Psychology, 1*(3), 154–165. https://doi.org/10.1080/21683603.2013.819306

Rowe, F. (2021, April 19). *Modelling count data in R: A multilevel framework*. https://fcorowe.github.io/countdata_modelling/

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Rudasill, K. M., & Rakes, C. R. (2008). *Jefferson County Public Schools Comprehensive School Surveys 2007-2008: Exploration and assessment of the structure of the surveys*. University of Louisville.

Rudasill, K. M., Snyder, K. E., Levinson, H., & Adelson, J. L. (2018). Systems view of school climate: A theoretical framework for research. *Educational Psychology Review, 30*, 35-60. https://doi.org/10.1007/s10648-017-9401-y

Ryberg, R., Her, S., Temkin, D., Madill, R., Kelley, C., Thompson, J., & Gabriel, A. (2020). Measuring school climate: Validating the Education Department School Climate Survey in a sample of urban middle and high school students. *AERA Open, 6*, 1-21. https://doi.org/10.1177/2332858420948024

Saint, J., Rice, K. G., Varjas, K., & Meyers, J. (2021). Teacher Perceptions Matter: Psychometric Properties of the Georgia School Personnel Survey of School Climate. *School Psychology Review, 50*(2-3), 406-419. https://doi.org/10.1080/2372966X.2021.1958645

Schweig, J., Hamilton, L. S., & Baker, G. (2019). *School and classroom climate measures: Considerations for use by state and local education leaders*. (RR-4259-FCIM). RAND Corporation. https://doi.org/10.7249/RR4259

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). AERA.

Shukla, K. D., Waasdorp, T. E., Lindstrom Johnson, S., Orozco Solis, M. G., Nguyen, A. J., Rodríguez, C. C., & Bradshaw, C. P. (2019). Does school climate mean the same thing in the United States as in Mexico? A focus on measurement invariance. *Journal of Psychoeducational Assessment, 37*, 55-68. https://doi.org/10.1177/0734282917731459

Steffgen, G., Recchia, S., & Viechtbauer, W. (2013). The link between school climate and violence in school: A meta-analytic review. *Aggression and Violent Behavior*, *18*(2), 300-309. https://doi.org/10.1016/j.avb.2012.12.001

Sun, L., & Royal, K. D. (2017). School climate in American secondary schools: A psychometric examination of PISA 2009 School Climate Scale. *Journal of Curriculum and Teaching, 6*(2), 6-12. https://doi.org/10.5430/jct.v6n2p6

Tatman, T. K. (2018, November 9). *Survey results show improved climate and culture within JCPS.* https://www.jefferson.kyschools.us/departments/communications/monday-memo/survey-results-show-improved-climate-and-culture-within-jcps#:~:text=88.6%20percent%20of%20respondents%20agreed,an%20increase%20of%204.3%20percent.

Tatman, T. K. (2019, September). *JCPS school board approves 2019-20 working budget.* https://www.jefferson.kyschools.us/departments/communications/monday-memo/jcps-school-board-approves-2019-20-working-budget

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83*(3), 357-385. https://doi.org/10.3102/0034654313483907

Thum, Y. M., & Kuhfeld, M. (2020). *NWEA 2020 MAP growth achievement status and growth norms for students and schools*. NWEA. https://teach.mapnwea.org/impl/normsResearchStudy.pdf

Turner, H. J., Natesan, P., & Henson, R. K. (2017). Performance evaluation of confidence intervals for ordinal coefficient alpha. *Journal of Modern Applied Statistical Methods, 16*(2), 157-185. https://doi.org/10.22237/jmasm/1509494940

U.S. Department of Education. (2014). *Guiding principles: A resource guide for improving school climate and discipline.* Washington, D.C. Retrieved from https://www2.ed.gov/policy/gen/guid/school-discipline/guiding-principles.pdf

U.S. Department of Education. (2020, December 11). *ED School Climate Surveys (EDSCLS)*. Retrieved from https://safesupportivelearning.ed.gov/edscls/edsclsvm4.4.7

Van Houtte, M., & Van Maele, D. (2011). The black box revelation: In search of conceptual clarity regarding climate and culture in school effectiveness research. *Oxford Review of Education, 37*(4), 505-524. https://doi.org/10.1080/03054985.2011.595552

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*, 53-62. https://doi.org/10.1093/cs/cdz026

Wang, M. T., & Degol, J. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Education Psychology Review 28*, 315-352. https://doi.org/10.1007/s10648-015-9319-1

Whitehouse, A., Zeng, S., Troeger, R., Cook, A., & Minami, T. (2021). Examining measurement invariance of a school climate survey across race and ethnicity. *Assessment for Effective Intervention, 47*(1), 37-46. https://doi.org/10.1177/1534508420966390

Yang, C., Bear, G. G., Chen, F. F., Zhang, W., Blank, J. C., & Huang, X. (2013). Students' perceptions of school climate in the U.S. and China. *School Psychology Quarterly, 28*(1), 7-24. https://doi.org/10.1037/spq0000002

Yang, C., Chan, M., Chen, C., & Jimerson, S. R. (2021). Parental perceptions of school climate in the United States and China: Advancing cross-country understanding. *School Psychology, 36*(1), 24-33. https://doi.org/10.1037/spq0000421

You, S., O'Malley, M. D., & Furlong, M. J. (2014). Preliminary development of the brief-California school climate survey: Dimensionality and measurement invariance across teachers and administrators. *School Effectiveness and School Improvement, 25*, 153-173. https://doi.org/10.1080/09243453.2013.784199

Zabek, F., Meyers, J., Rice, K. G., Ashby, J. S., & Kruger, A. C. (2022). Can a school climate survey accurately and equitably measure school quality? Examining the multilevel structure and invariance of the Georgia School Climate Scale. *Journal of School Psychology*, *95*, 1-24. https://doi.org/10.1016/j.jsp.2022.08.005

Zullig, K. J., Collins, R., Ghani, N., Hunter, A. A., Patton, J. M., Huebner, E. S., & Zhang, J. (2015). Preliminary development of a revised version of the School Climate Measure. *Psychological Assessment, 27*(3), 1072-1081. http://dx.doi.org/10.1037/pas0000070

Zullig, K. J., Collins, R., Ghani, N., Patton, J. M., Scott Huebner, E., & Ajamie, J. (2014). Psychometric support of the school climate measure in a large, diverse sample of adolescents: A replication and extension. *Journal of School Health, 84*, 82-90. https://doi.org/10.1111/josh.12124

Zullig, K. J., Huebner, E. S., & Patton, J. M. (2011). Relationships among school climate domains and school satisfaction. *Psychology in the Schools, 48*(2), 133–145. https://doi.org/10.1002/pits.20532

Zullig, K. J., Koopman, T. M., Patton, J. M., & Ubbes, V. A. (2010). School climate: Historical review, instrument development, and school assessment. *Journal of Psychoeducational Assessment, 28*, 139-152. https://doi.org/10.1177/0734282909344205

CURRICULUM VITA

NAME:	Stephen M. Leach
EDUCATION & TRAINING:	Ph.D., Educational Psychology Measurement & Evaluation University of Louisville, 2022
	B.A., Mathematics University of Louisville, 2010
PROFESSIONAL EXPERIENCE:	*Grant Developer*, 2022-Present Jefferson County Public Schools, Louisville, KY
	Program Analysis Coordinator, 2018-2022 Jefferson County Public Schools, Louisville, KY
PUBLICATIONS:	*Peer-Reviewed Journals*
	Hollands, F.M., Leach, S.M., Shand, R., Head, L., Wang, Y., Dossett, D., Chang, F., Yan, B., Martin, M., Pan, Y. (2022). Restorative Practices: Using local evidence on costs and student outcomes to inform school district decisions about behavioral interventions. *Journal of School Psychology, 92*, 188-208. https://doi.org/10.1016/j.jsp.2022.03.007
	Hollands, F.M., Shand, R. Yan, B., Leach, S.M., Dossett, D., Chang, F., & Pan, Y. (2022). A comparison of three methods for providing local evidence to inform school and district budget decisions. *Leadership and Policy in Schools*. https://doi.org/10.1080/15700763.2022.2131581
	Leach, S.M., Hollands, F.M., Stone, E., Shand, R., Head, L., Wang, Y., Yan, B., Dossett, D., Chang, F., Chang, Y., & Pan, Y. (2022). Costs and effects of school-based licensed practical nurses on elementary student attendance and chronic absenteeism. *Prevention Science*. https://doi.org/10.1007/s11121-022-01459-0
	Mitchell, A.M., Heitz, H.K., Leach, S.M., Berghuis, K.J., (2022). Material circumstances, health care access, and self-reported health: A latent class analysis. *Journal of Health Psychology*. https://doi.org/10.1177/13591053221132899

Shand, R., Leach, S.M., Hollands, F., Chang, F., Pan, Y., Yan, B., Dossett, D., Nayyer-Qureshi, S., Wang, Y., & Head, L. (2022). Program value-added: A feasible method for providing evidence on the effectiveness of multiple programs implemented simultaneously in schools. *American Journal of Evaluation*. https://doi.org/10.1177/10982140211071017

Ingle, W.K., Leach, S.M., Lingo, A.S. (2021). Assessing efforts to diversify Kentucky's K-12 teacher workforce: A mixed-methods analysis of a grow-your-own teacher pathway. *Journal of Education Human Resources*. Advance online publication. https://doi.org/10.3138/jehr-2021-0038

Leach, S.M., Mitchell, A.M., Salmon, P., & Sephton, S.E. (2021). Mindfulness, self-reported health, and cortisol: A latent profile analysis. *Journal of Health Psychology, 26*(14), 2719-2729. https://doi.org/10.1177/1359105320931184

Leach, S.M., Immekus, J.C., French, B.F., & Hand, B. (2020). The factorial validity of the Cornell Critical Thinking Tests: A multi-analytic approach. *Thinking Skills and Creativity, 37*, 100676. https://doi.org/10.1016/j.tsc.2020.100676

Choi, N., Leach, S.M., Hart, J.M., & Woo, H. (2019). Further validation of the Brief Resilience Scale from a Korean college sample. *Journal of Asia Pacific Counseling, 9*, 39-56. https://doi.org/10.18401/2019.9.2.3

Valentine, J.C., Leach, S.M., Fowler, A., Stojda, D.K., & Macdonald, G. (2019). Families and Schools Together (FAST) for improving outcomes for children and their families. *Cochrane Database of Systematic Reviews, 7*, Article No. CD012760. https://doi.org/10.1002/14651858.CD012760.pub2

Research Briefs and Other Writings

Leach, S.M., Shand, R., Yan, B., & Hollands, F. (2022, February 15). *Unexpected value from conducting value-added analysis*. IES Blog. https://ies.ed.gov/blogs/research/post/ unexpected-value-from-conducting-value-added-analysis

Leach, S.M., Hollands, F., Yan, B., & Shand, R. (2022, February 3). *Unexpected benefits of conducting cost-effectiveness analysis*. IES Blog. https://ies.ed.gov/blogs/research/post/ unexpected-benefits-of-conducting-cost-effectiveness-analysis

Leach, S.M. & Yan, B. (2021, December). *Academic return-on-investment (AROI) and budget decision-making: A research brief.* Louisville, KY: Jefferson County Public Schools https://www.jefferson.kyschools.us/sites/default/files/IES_AROI_Brief_3_Final_Dec_2021.pdf

PRESENTATIONS & CONFERENCE PAPERS:

Shand, R., Leach, S., Yan, B., Hollands, F., Dossett, D., Chang, F., & Pan, Y. (2022, September). *Rethinking success: The unexpected benefits of research-practice partnerships.* Accepted to the SREE 2022 Conference, Washington, D.C.

Hollands, F.M., Leach, S.M., Stone, E., Head, L., Wang, Y., Shand, R.L, Yan, B., Dossett, D.H., Chang, F., Chang, Y., Pan, Y. (2021, April). *Costs and effects of school nursing on student attendance and absenteeism*. Accepted to the American Education Research Association (AERA) Annual Conference, Virtual.

Hollands, F., Shand, R., Yan, B., Leach, S., Dossett, D., Chang, F., & Pan, Y. (2021, March). *A comparison of three methods for providing local evidence to inform school district budget decisions*. Presented at the AEFP Annual Conference, Virtual.

Hollands, F., Leach, S., Shand, R., Head, L., Wang, K., Dossett, D., Chang, F., Yan, B., & Pan, Y. (2021, March). *Restorative Practices: Using local evidence on costs and student outcomes to inform cycle-based budget decisions*. Presented at the AEFP Annual Conference, Virtual.

Leach, S.M., Ingle, W.K., & Lingo, A. (2020, April). *Assessing efforts to diversify Kentucky's K-12 teacher workforce: A mixed-methods approach.* Accepted to the AERA Annual Conference, San Francisco, CA.

Shand, R., Pan, Y., Leach, S.M., Nayyer-Qureshi, S., Hollands, F., Yan, B., Dossett, D., Wang, Y., Head, & L., Chang, F. (2020, March). *Program evaluation with administrative data: New applications, promise, and challenges for value-added models.* Accepted to the Association for Education Finance and Policy (AEFP) Annual Conference, Fort Worth, TX.

Hollands, F., Yan, B., Leach, S.M., Shand, R., Dossett, D., Wang, Y., & Head, L. (2020, March). *LEA use of evidence in budget decisions.* Accepted to the SREE 2020 Conference, Arlington, VA.

Hollands, F. Leach, S.M., Feng, M., & Reichardt, R. (2020, January). *Cost-effectiveness analysis for research. Cost-effectiveness analysis for practice.* Presented at the Institute for Educational Sciences (IES) Principal Investigators (PI) Meeting, Washington, D.C.

Shand, R., Leach, S.M., Nayyer-Qureshi, S., & Pan, Y. (2020, January). *Applying value-added analysis to program evaluation.* Accepted to the IES PI Meeting, Washington, D.C.

Leach, S.M., Snyder, K.E., Potter, D.A., & Immekus, J.C. (2019, November). *A Bioecological Metatheory approach to evaluability assessment*. Presented at the American Evaluation Association Annual Conference, Minneapolis, MN.

Dossett, D. Yan, B., & Leach, S.M. (2019, October). *Track investments to improve strategic use of financial resources*. Presented at the Annual Fall Conference of Council of Great City Schools, Louisville, KY.

Fowler, A., & Leach, S.M. (2019, September). *A gentle introduction to multilevel modeling*. Presented to faculty and graduate students at the University of Louisville College of Education and Human Development, Louisville, KY.

Dossett, D., Yan, B., & Leach, S.M. (2019, May) *Can investment tracking improve strategic budgeting?* Presented at Strategic Data Project Convening 2019, Harvard University Center for Education Policy Research, Boston, MA.

Leach, S.M., Immekus, J.C., & French, B.F. (2019, April). *The factorial validity of the Cornell Critical Thinking Tests: A multi-analytic approach.* Accepted to the AERA Annual Conference, Toronto, CA.

Choi, N, Leach, S.M., Hart, J.M., & Woo, H. (2017, April). *Validation of the Brief Resilience Scale from a Korean college sample*. Accepted to the AERA Annual Conference, San Antonio, TX.

Leach, S.M., Samuelsson, M., & Fiet, J.O. (2017). The paradox of effectual search. *Frontiers of Entrepreneurship Research, 37*(16). Article 9.

PROFESSIONAL AFFILIATIONS & SERVICE:

Professional Affiliations

American Evaluation Association, 2018-Present

American Educational Research Association, 2017-Present

Grant Professionals Association, 2022-Present

Service

Invited Reviewer, *Use of Research Evidence Across Settings Section*, SREE 2022 Conference

Invited Reviewer, *Research Methods Section,* SREE 2021 Conference

Managing Editor, *Politics of Education Book Series*, Information Age Publishing

Ad Hoc Reviewer, *Journal of Education Human Resources*

Ad Hoc Reviewer, *Measurement and Evaluation in Counseling and Psychology*

Ad Hoc Reviewer, *Thinking Skills and Creativity*

YMCA Volunteer Youth Soccer Coach, (2021-2022)

AWARDS:

J.H. Morris Scholarship, 2020

Doris J. Bouse Trautwein Scholarship, 2020

CEHD Summer Scholarship, 2019

CEHD Spring Scholarship, 2019

Research and Faculty Development Grant, 2018

CEHD Winter Scholarship, 2018