Improving the Measurement of School Climate Using Item Response Theory

Sarah Lindstrom Johnson[a], PhD (corresponding author); Ray E. Reichenberg, MEd[a], MA;

Kathan Shukla[b], PhD Tracy E. Waasdorp[c], PhD; Catherine P. Bradshaw[d], PhD

[a]Arizona State University
Social Sciences Room 116
PO Box 873701, Tempe AZ 85287-3701
(480) 965-9975 Fax: (480) 965-6779
Sarahlj@asu.edu; rreiche@asu.edu

[b]Indian Institute of Management, Ahmedabad
Vastrapur, Ahmedabud 380015
Gujarat, India
kathans@iiama.ac.in

[c]Johns Hopkins School of Public Health
415 N. Washington Street,
Baltimore, MD 21231
twaasdo1@jhu.edu

[d]University of Virginia
112-D Bavaro Hall, 417 Emmet Street South, PO Box 400260
Charlottesville, VA 22904-4260
Catherine.bradshaw@virginia.edu

Johnshon, S.L., 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). DOI: https://doi.org/10.1111/emip.12296

This is a post-peer-review, pre-copyedit version of an article published in Educational Measurement Issues and Practice. The final authenticated version is available online at: https://onlinelibrary.wiley.com/doi/full/10.1111/emip.12296

Acknowledgement: This work was funded in part by grants from the U.S. Department of Education to the Maryland State Department of Education and the Institute for Educational Sciences (R305H150027) and the National Institute of Justice (2014-CK-BX-0005) to Catherine Bradshaw.

Abstract

The United States government has become increasingly focused on school climate, as recently evidenced by its inclusion as an accountability indicator in the *Every Student Succeeds Act*. Yet, there remains considerable variability in both conceptualizing and measuring school climate. To better inform the research and practice related to school climate and its measurement, we leveraged item-response theory (IRT), a commonly used psychometric approach for the design of achievement assessments, to create a parsimonious measure of school climate that operates across varying individual characteristics. Students (*n*= 69,513) in 111 secondary schools completed a school climate assessment focused on three domains of climate (i.e., safety, engagement, environment), as defined by the U.S. Department of Education. Item and test characteristics were estimated using the 'mirt' package in *R* using unidimensional item response theory. Analyses revealed measurement difficulties that resulted in a greater ability to assess less favorable perspectives on school climate. Differential item functioning analyses indicated measurement differences based on student academic success. These findings support the development of a broad measure of school climate but also highlight the importance of work to ensure precision in measuring school climate, particularly when considering use as an accountability measure.

Keywords: item response theory; school climate; school accountability; measurement invariance

School climate has been defined as the quality and character of school life and relates to norms, values, and expectations that foster supportive environments and feelings of engagement and safety. A favorable school climate has been associated with both improved behavioral and academic outcomes for students including increased academic achievement and reduced suspension, absenteeism, truancy, dropout, drug use, and violent and aggressive behavior (for a review, see Thapa, Cohen, Guffey, & Higgins-D'Alessandro, 2013). School climate is conceptualized as a property of the school, but traditionally assessed through individuals' perceptions (Bradshaw, Waasdorp, Debnam, & Lindstrom Johnson, 2014). These perceptions, both at the individual level and at the collective level, are shaped by internal and external factors as well as by shared experiences of school life (Cohen, McCabe, Michelli, & Pickeral, 2009). Yet important debates still remain regarding the definition of school climate and its measurement, which have implications for determining the causal link between climate, academics, and behavioral outcomes (Payne, 2018).

The United States federal government has become increasingly focused on school climate. As of the 2017-2018 school year, the *Every Student Succeeds Act* (ESSA) requires that states, along with traditional indicators of academics (e.g., graduation, proficiency in reading and math), include one other indicator of school quality or success, such as student engagement, safety, or school climate in their accountability indicator. The majority of states have chosen to use archival data (e.g., chronic absenteeism, access to advanced coursework or career exploration offerings) to meet the reporting requirement (Achieve, 2019), and relatively few states have opted to administer a survey of students' perceptions. This decision likely reflects, at least in part, the current state of the field of school climate research regarding uncertainty surrounding its

conceptualization as well as its measurement. The intent of this paper is to inform the measurement of school climate and its possible use as an accountability indicator.

Defining and Measuring School Climate

Conceptualizations of school climate can vary in breadth and specificity (Lindstrom Johnson, Waasdorp, & Bradshaw, in press), which has a direct implication for what is measured and reported. As such, there remains a considerable conceptual debate regarding the parameters of school climate. Specifically, some recent work suggests that school climate might best be captured by a narrower measure of student engagement (Payne, 2018). Other related work has defined school climate as "the quality and consistency of interpersonal interactions within the school community that influence children's cognitive, social and psychological development" (Haynes, Emmons, & Ben-Avie, 1997, p, 322). This conceptualization posits that the constructs of school safety and aspects of the school environment are determinants of school climate, or outcomes of school climate, but they are not part of the specific measure of school climate. This distinction reflects a broader debate regarding whether the focus should be on school safety as compared to school climate (Benbenishty, Astor, Roziner, & Wrabel, 2016).

Nevertheless, the U.S. Department of Education (USDOE) put forth an inclusive model of school climate which reflects both student safety and the school environment, as well as student engagement. Specifically, *engagement* focuses on relationships between students, staff, and families which are built on trust and respect and foster connection with the school. These features of school climate are thought to be fostered by a school *environment* with clear rules and expectations and supports for learning. Engagement and environment create and are supported by physical and emotional *safety* (National Center on Safe Supportive Learning Environments, 2018). Part of the reason for this broad conceptualization may reflect a desire to encourage

consideration of all of these constructs, which individually have been linked with student learning and behavioral outcomes (Lindstrom Johnson et al., in press). However, the use of a multidimensional model of school climate presents difficulties in understanding how to aggregate this data into a single accountability measure. In fact, the desire for a school-level indicator of school climate for school-level accountability was a motivating factor behind the USDOE's Safe and Supportive Schools grant; this multi-million federal dollar initiative funded 11 states to develop a comprehensive measure of school climate and pilot the measures in high schools to inform the implementation of evidence-based programs to improve school climate (Bradshaw et al., 2012; Shaw, 2013).

Individual Variability in Perceptions of School Climate

While there is substantial variability at the classroom and school-level, the majority of variability in perceptions of school climate is attributable to individual differences (Fan, Williams, & Corkin, 2011; Koth, Bradshaw, & Leaf, 2008). Recent efforts to validate surveys of school climate have provided evidence of measurement invariance across a range of demographic factors (e.g., gender, age, and race) (Bear, Gaskins, Blank, & Chen, 2011; Bradshaw et al., 2014). These findings give confidence that school climate can be conceptualized the same across groups, and that discrepancies that do occur are meaningful and not solely attributable to differences in measurement quality. For example, research has found that girls are more likely to report a positive school climate (Kuperminc, Leadbeater, Emmons, & Blatt, 1997) including a higher achievement motivation (Koth et al., 2008) and better relationships with teachers (Crosnoe, Johnson, & Elder, 2004), whereas boys are more likely to report lower levels of order and discipline (Koth et al., 2008) and disciplinary problems (Crosnoe et al., 2004). Youth of color tend to report less supportive relationships with their teachers, have lower

perceptions of equity, and perceive the environment as less safe (Bottiani, Bradshaw, & Mendelson, 2016; Fan et al., 2011). School climate perceptions have been shown to decrease through the transition from elementary to middle school (Espinoza & Juvonen, 2011) and across middle school (Way, Reddy, & Rhodes, 2007) and improve throughout the course of high school (Bradshaw et al., 2014; Crosnoe et al., 2004). The influence of school climate on outcomes has also been shown to vary (i.e., be moderated) by gender (Crosnoe et al., 2004; Henry et al., 2011; Kuperminc et al., 1997), race (Crosnoe et al., 2004; Espinoza & Juvonen, 2011; Kuperminc et al., 1997), and age (Henry et al., 2011).

Less empirical research has explicitly focused on differences in perceptions by levels of academic success or parental education. A study examining latent profiles of student perceptions of school climate found that students who perceived a more positive climate had higher mean academic outcomes and came from families with higher levels of parental education than those who perceived a negative climate (Shukla, Konold, & Cornell, 2016). Students who perform better academically may have different experiences with teachers, or be more likely to feel engaged to school (Battish, Solomon, Kim, Watson, & Schaps, 1995; Kuperminc et al., 1997). Further, Fan and colleagues (2011) found evidence of an association between parental educational level and perceptions of order, safety, and discipline but not teacher-student relationships or fairness and clarity of school rules.

Application of Item Response Theory to the Assessment of School Climate

An innovative aspect of the current paper was the use of IRT analyses to examine a measure of school climate that is aligned with the USDOE's conceptualization, with the overarching goal of creating a more parsimonious, yet psychometrically sound measure of climate. While IRT has been previously applied to school climate scales (see Mo, Yang, & Hu,

2011), we capitalized on the existence of measures across three possible domains of school climate (e.g., safety, engagement, environment) to both understand the ability of diverse items to create a scale, as well as explore how the varying domain scales operated. Specifically, we assessed differential item functioning for each scale in relation to a broad range of individual demographic characteristics, including maternal education (a measure of socio-economic status) and grades (a measure of academic success). We also evaluated the ability of each scale to assess the continuum of perceptions of school climate. Taken together, the result of the IRT analyses are intended to further illustrate the validity, reliability, and potential usability of this particular measure of school climate for schools and state education agencies, as well as educational researchers. As such, the current study had a dual focus on informing both the measurement and conceptualization of school climate, which ultimately may inform schools' and states' use of surveys to meet expectations outlined in *ESSA*.

Method

Initial Data Collection

Procedures. Data for the study came from the Maryland Safe and Supportive Schools (MDS3) Initiative, which is a collaborative effort of the Maryland Department of Education (MSDE), Johns Hopkins University, and Sheppard Pratt Health System aimed at improving school climate and student outcomes. The MDS3 School Climate Student Survey (Bradshaw et al., 2014) was developed as a self-report survey and is delivered on-line to students, staff, and parents in public middle and high schools across the state of Maryland. In the current study, we drew upon data from 111 schools across 13 Maryland school districts. Districts were approached for participation by the MSDE. Upon expressing interest in the MDS3 Initiative, district-specific principal meetings were conducted to obtain school-level and principal commitment to the

project. The anonymous survey was administered using a passive consent and youth assent process, and all participation was voluntary. Letters were sent home to parents providing information about the survey and the larger initiative. The survey was administered online in language arts classrooms at participating high schools. School staff provided instructions for students to complete the survey following a written protocol developed by the research team. The non-identifiable data were obtained from MSDE for analysis for the current paper. The non-identifiable data analysis was approved by researchers' Institutional Review Boards.

Participants. Data come from 69,513 secondary school students; 46% of students were in the 6th, 7th, or 8th grade (i.e., middle school) with the remainder in high school. Approximately half of the students identified as male. The sample was fairly diverse with 48.8% of students identifying as White, 25.7% as Black, and 9.6% as Hispanic. Almost 30% of students reported their mothers had a high school education or less with 55% reporting their mothers had obtained a high school degree. A majority of students (79.2%) self-reported earning mostly A's or B's on their last report card. Further description of participant demographics and school demographics can be seen in Table 1.

Insert Table 1 Here

Instrument Design

The MDS3 School Climate Student Survey (Bradshaw et al., 2014) was developed by the Johns Hopkins Center for Youth Violence Prevention in collaboration with project partners. Researchers from the Center undertook a comprehensive review of the literature focusing on the three domains of school climate included in the USDOE (2009) model (i.e., safety, engagement, and environment). Additionally, focus groups were held with students, district personnel, and school administrators to understand the operationalization of school contextual factors for each

of the different stakeholders. The full survey contains over 150 items and takes approximately 20 minutes for students to complete online.

The initial measurement model was created using an exploratory and confirmatory factor analytic approach in a high school sample (Bradshaw et al., 2014). Fifty-six core items based on previously validated domains of safety, engagement, and the school environment were identified. Each domain contains multiple scales: *Safety* includes the scales perceived safety (4 items; α= .64), bullying and aggression (4 items; α= .63), and general drug use (3 items; α= .87), *Engagement* includes the factors connection to teachers (6 items; α= .86), student connectedness (5 items; α= .87), academic engagement (4 items; α= .79), whole school connectedness (4 items; α= .82), culture of equity (4 items; α= .83), parent engagement (4 items; α= .74), and *Environment* includes the factors rules and consequences (5 items; α= .73), physical comfort (4 items; α= .79), support (4 items; α= .76), and disorder (5 items; α= .58). All answer choices were on a 4-point Likert scale from strongly agree to strongly disagree (unless otherwise noted), whereby all items were coded with high score representing a more favorable school climate. A specific research question for this paper was the extent to which a subset of items from each domain could create a domain-scale (i.e., Safety, Engagement, Environment).

IRT analyses explored measurement invariance across groups by gender (male/female), race (minority/non-minority), and grade in school (upper/lower classman). These analyses found evidence of scalar invariance (Bradshaw et al., 2014). Additional analyses also found measurement invariance across middle and high school students' reports (Waasdorp, Shukla, Lindstrom Johnson, & Bradshaw, under review). Other studies have compared the functioning of scales to observations of the school social and physical environment (Bradshaw, Milam, Furr-Holden, & Lindstrom Johnson, 2015). However, to date, there have been no efforts to address

parsimony, assess differential functioning by academic success and socioeconomic status, and explore item and scale functioning across the continuum of perceptions of climate. Comparisons of these findings across the three scales were also examined.

Overview of the Item Analysis Procedures

Using the 'mirt' package (Chalmers, 2012) in R (R Development Core Team, 2008), item characteristics were estimated for each of the three domains (e.g., safety, engagement, and environment) through unidimensional IRT using a graded response model (Samejima, 1997). These estimates included item location, item discrimination, and item information. The decision to estimate unidimensional models was based on a) computational ease, and b) the lack of any hypothesized cross-loadings (i.e., multidimensional models would have been specified to exhibit simple structure). The item and test information estimates resulting from the fitting of the IRT models as well as a selection of local and global fit statistics (discussed below) were then used to further refine the scales with the goal of creating the shortest scales possible while retaining a breadth of constructs and an adequate amount of test information across the spectrum of participant perspectives (i.e., three standard deviations below to three standard deviations above a neutral perspective on school climate). Adequacy was determined by converting test information to reliability using the equation put forth by Thissen (2000) and targeting a reliability estimate greater than 0.70 across the spectrum. Global fit statistics of final models were assessed including root mean squared error of approximation (RMSEA), comparative fit index (CFI), the squared root mean standardized residuals (SRMSR), and the M^2 statistic (see Hu & Bentler, 1999 for the first three and Maydeu-Olivares & Joe, 2006 for the last). Reliability statistics were also assessed including McDonald's Omega total (ω_t; McDonald, 1999), and Cronbach's α (Cortina, 1993).

Differential item functioning (DIF; Holland & Wainer, 2012) analyses were conducted in order to identify any items exhibiting bias across any of five areas: gender, minority status, academic success, status as a middle school or high school student, or level of maternal education. Each of these variables were dichotomized with the resulting designations being [male; female], [White, non-White], [A/B, C or worse], [middle school, high school], [less than a high school education, high school education or greater], respectively. Typical methods of detecting DIF often involve conducting likelihood ratio tests of nested models wherein one model constrains the item parameters to be equal across groups and a comparison model allows those parameters to vary. These methods were not feasible in this case given the large sample size as any such test would be drastically overpowered and, therefore, overly sensitive to potentially spurious group differences. To address this potential concern, effect sizes (see Meade, 2010) were used in lieu of significance tests following the method of Meade and Wright (2012). The use of such effect sizes necessitated defining criteria for what constitutes an acceptable difference as no criteria or recommendations currently exist in the DIF literature. For the purposes of the current study, it was decided that typical effect size interpretations would be used for standardized metrics and that an expected focal group difference, relative to the reference group, of one scale point at the test-level would be utilized for unstandardized metrics. Finally, factor scores (theta values) were estimated using the expected a posteriori (EAP) method (Embretson & Reise, 2000). Mean differences in distributions of theta by the above group differences were also explored using Hedges' *g* as a measure of effect sizes (Hedges, 1981).

Results

Measure Creation

A primary goal of the scale construction process was to create a scale that provided a reliable estimate of a respondent's perspectives of school climate within each of the three domains (e.g., safety, engagement, environment). To this end, item and test information curves were examined and used in the scale refinement process. A total of 30 of the 56 core items were retained across the three scales (10 items per scale). Table 2 presents descriptive statistics as well as the IRT parameter estimates for the items included in the final scales.

Insert Table 2 Here

Table 2 presents descriptive statistics as well as IRT parameter estimates for the items included in the final version of the scales. The discrimination parameter (a) refers to an item's ability to differentiate between respondents of different trait levels while the threshold parameters (b_1, b_2, b_3) indicate the difficulty of the item. In the case of the graded response model, these thresholds indicate the points at which a respondent would have the same probability of endorsing any of the categories below the threshold compared as they would for the categories above the threshold. In the case of the first item on the *Safety* scale ("I feel safe at this school."), for example, the probability of a respondent 1.01 standard deviations above the mean on the latent *Safety* trait choosing "Strongly Agree" (i.e., $P(x = 3) = 0.50$) would be the same as their probability of choosing any of the other responses (i.e., $P(x = 0 \text{ or } x = 1 \text{ or } x = 2) = 0.50$). As can be seen in Table 2, the discrimination parameter estimates (a) for the majority of items in the *Safety* scale and some in the *Environment* scale were fairly low. As the discrimination parameter assesses the ability of an item to differentiate between two individuals, low values indicate items that provide limited information. From a psychometric perspective, an ideal scale would include a mix of items that provide moderate information across the scale as well as items that provide high information at varying ability levels.

Figure 1 presents the final test information curves for each of the three scales of *Safety*, *Engagement*, and *Environment*. For convenience, horizontal lines have been drawn on the test information curves at the points along the Y-axis corresponding to reliability estimates of 0.70, 0.80, and 0.90.

Insert Figure 1 Here

Figure 1 presents the test information curves for each of the three scales. These curves indicate the information that the scale is able to provide as well as the reliability of the test for a given level of the latent trait. The shape of the curves suggests that the items in the scales tended to provide more information (and, thus are more reliable) for respondents with less favorable perceptions of school climate. This was particularly true for the *Safety* scale for which estimates of scale reliability drop below acceptable levels at perceptions 1 SD above mean (e.g., see Figure 1 where the reliability drops below .8 at 1SD and below .7 at 2 SDs above). Upon examination of the threshold values b_i in Table 2 (which refer to the points at which the probability of choosing response $i + 1$ or higher is equal to .5), it can be seen that most items were located on the left-side of the school climate spectrum suggesting that these items might be better at capturing negative perceptions of school climate than positive.

For model fit, the values of the RMSEA, CFI, and SRMR indices generally met the criteria for satisfactory global fit set forth in the literature with a few exceptions (see Table 3).

Insert Table 3 Here

The CFI value for the *Environment* model (0.867) fell below the commonly applied criteria of 0.90 and the SRMR value for the *Safety* model was larger than the suggested cutoff of 0.08. Neither of these departures was deemed to be reason enough for discarding the respective models. All reliability statistics were acceptable with the exception of McDonald's (1999)

hierarchical Omega (ω_h) for Safety. Specifically, the *Safety* scale seemed to have less general commonality and more group commonality. This suggests the existence of a subset of items in the *Safety* scale that were strongly related beyond the general factor. The other two scales of *Environment* and *Engagement* had stronger general factors and less variance associated with subsets of items. Additional details about global model fit and reliability statistics for each of the three unidimensional models can be found in Table 3. Correlations between the three factors were moderate to high (*Safety* with *Engagement*, $r = .619$, $p<.05$; *Safety* with *Environment*, $r = .656$, $p<.05$; *Engagement* with *Environment*, $r = .853$, $p <.05$).

Individual Differences in Perceptions of School Climate

Results from the DIF analyses are presented in Table 4. Included in this table are two effect size estimates; the signed test difference in the sample (STDS) and the expected test score standardized difference (ETSSD). Note that the STDS uses signed estimates of group difference and, as such represents the difference in expected scale scores averaged across all focal group respondents (Meade, 2010). The STDS estimate is interpreted in the metric of the scale of the instrument (e.g., 0-30 for a 10-item instrument with each item being scored [0,1,2,3]). The ETSSD estimate is a standardized effect size; the same rules that apply to interpreting other effect sizes can be used to interpret these values. The results indicate that there are no substantial measurement differences by gender, minority status, academic success, status as a middle school or high school student, or level of maternal education. The largest differences are between students of differing academic status, with students reporting getting A's and B's on their report card on average scoring .595 points higher on the *Safety* scale. However, this is still considered a small effect (i.e., ETSSD of -.112). Overall, the *Environment* scale exhibited the lowest level of differences between subgroups, with the possible exception of minority status.

Insert Table 4 Here

Factor scores (i.e., theta values) were plotted for each of the subgroups; Table 4 presents effect sizes of difference. Most are considered small, with some differences by students' self-reported grade in school being slightly more substantial. Examining plots of factor scores by subgroups shows a positively shifted distribution (see Figure 2), whereby students who received higher grades had more favorable perceptions of safety, engagement, and environment. To a lesser degree, students whose parents had more than a high school degree also held more favorable perceptions of safety, engagement, and environment.

Discussion

This paper is novel in its application of IRT to improving the measurement of students' perceptions of school climate using the USDOE's model. Using this method, we created parsimonious scales measuring the three key USDOE (2009) domains of school climate (i.e., safety, engagement, environment). This was done with specific attention to ensure that items in the scales covered the breadth of topics from component subscales. Additionally, we demonstrated that the scales equivalently assessed safety, engagement, and environment across gender, minority status, academic success, status as a junior or high school student, and maternal education, but that differences in perceptions, particularly by academic success, existed. Attention to these measurement differences is important as individual evaluations of school climate are often aggregated to the school level as a metric of school performance (Battish et al., 1995; Way et al., 2007).

Through the process of evaluating item functioning and then scale composition, we identified that our measure was better able to assess differences in individual's perceptions of poor school climate than of excellent school climate. This was most apparent in the test

information curves (see Figure 1) that show more information below the average theta than above. This is primarily the result of "easy" items, in which the probability of getting an answer "correct" is fairly high. Take for example the item "Students carrying guns or knives at this school is a problem". At around 2 *SD*s below the mean students have an equal probability of saying they *strongly agree* with this statement versus any other response. At .5 *SD* still below the mean they have an equal probability of saying they *strongly disagree* with this statement versus any other response. While all three scales provided more information at the less favorable end of perceptions, this was particularly problematic for assessments of school safety. This may be the artifact of students' feeling safe at school (Musu-Gillette et al., 2018). However, it does have implications for measurement accuracy, as it suggests that we can more reliability measure the perceptions of those who perceive their school to have poor school climate than excellent school climate, and may support the need for a more nuanced understand of safety that goes beyond acts of violence and perceptions of safety (Embretson & Reise, 2000). It also has implications for school climate interventions as it suggests that schools with more positive school climates may be less invested in school climate assessments and interventions (Bradshaw et al., 2014).

The current study is one of the first to examine measurement invariance for students reporting different levels of academic engagement as well as differing levels of socioeconomic status (Bear et al., 2011; Bradshaw et al., 2014). Our results suggested that differences resulting from the models across these domains as well as gender, minority status, and level of school could be attributed as meaningful and not as the result of error. Additionally, effect size DIF estimates suggested limited mean differences in perceptions of school climate across groups at the same level of perceptions of school climate (i.e., theta). This is a potentially novel contribution to the understanding of individual differences in school climate perceptions as it

takes into account group difference in perceptions of climate (i.e., items function the same for girls who view the school climate as less favorable as boys who view the climate as less favorable). Although most group differences in theta were small, there were moderate differences for students who had higher grades. Our results suggest that schools with a higher percentage of students receiving A's and B's will receive higher aggregated school climate scores simply as a reflection of measurement difference; therefore, schools with a higher concentration of better performing students on average tend to rate their school climate more favorably. This suggests the difficulty of disentangling school climate and achievement and may explain differential findings around the relationship between school climate and achievement, particularly those that involve the aggregation of student perceptions (Benbenishty et al., 2016).

An additional contribution of this paper is its potential to inform discussions regarding the definition of school climate. This is important as it has direct implications for what is measured and used as an accountability measure. Due to the multidimensional nature of school climate, surveys of student perceptions can be lengthy which may dissuade state's from using them. Furthermore, it is difficult to take the data from a multi-dimensional construct to create a single accountability indicator. Together, our results suggest the possibility of measuring school climate across the broad categories of safety, engagement, and environment. More work is needed to understand how these three indicators might be aggregated (Bradshaw et al., 2014) or how they might be causally related (Payne, 2018). Nevertheless, our findings suggested that engagement and environment are more highly correlated than safety, which may reflect the conceptual challenges in defining safety.

Strengths, Limitations, and Next Steps

A strength of this study is that we started with a previously validated measure of school climate which is consistent with the USDOE's conceptualization of school climate. We aimed to create a more parsimonious measure that would be both more time efficient for schools to use but also psychometrically strong from an IRT perspective. Future work should also begin to address the likely correlations between the various components of school climate (i.e., safety, engagement, and environment), allowing for a possible aggregate measure of school climate. Additional work should also focus on aspects of external validity by determining the extent to which the scale scores relate to student behavioral indicators of interest to educators and policy makers, such as suspensions, academic performance, and high school completion. Moreover, future studies could also contrast the concurrent and predictive validity of the long vs. short version in reference to these and other student behavioral indicators to insure that the shortened version is in fact sufficiently predictive of particular outcomes of interest. An important limitation to note is that the data were drawn from one state, and middle and high schools and students. Staff and even parents may provide additional valuable insights into school climate (Waasdorp, Pas, O'Brennan, & Bradshaw, 2011).

Conclusion

This paper applied sophisticated analytic techniques used for assessment design to measures of school climate, with the overarching goal of making a measure, which is consistent with the USDOE's model of school climate, more efficient while not compromising its validity and reliability. Leveraging advanced psychometric tools, which have largely focused on academic and other measurement topics, we helped to advance the field of school climate assessment. In doing so we also contribute to debates about the scope and conceptualization of school climate. This line of work is particularly timely in light of *ESSA's* emphasis on school

climate and related constructs. Furthermore, our findings highlight the importance of disentangling student background from school variables. Although a common language is emerging regarding the various theorized dimensions of school climate, the findings of this study advance the conversation by providing insight on how to both efficiently and precisely measure these three core dimensions of school climate. This type of empirical work is critical to support the inclusion of a broader array of school factors into discussions of accountability for school leaders.

References

Achieve (2019). *Accountability in state ESSA plans.* Retrieved from https://states.achieve.org/essa-tracker

Anderson, C. S. (1982). The search for school climate: A review of the research. *Review of Educational Research, 52*, 368-420.

Battistich, V., Solomon, D., Kim, D. I., Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American Educational Research Journal*, *32*(3), 627-658.

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.

Benbenishty, R., Astor, R. A., Roziner, I., & Wrabel, S. L. (2016). Testing the causal links between school climate, school violence, and school academic performance: A cross-lagged panel autoregressive model. *Educational Researcher*, *45*(3), 197-206.

Bottiani, J. H., Bradshaw, C. P., & Mendelson, T. (2016). Inequality in black and white high school students' perceptions of school support: An examination of race in context. *Journal of Youth and Adolescence*, *45*(6), 1176-1191.

Bradshaw, C. P., Milam, A. J., Furr-Holden, C. D. M., & Lindstrom Johnson, S. (2015). The School Assessment for Environmental Typology (SAfETy): An observational measure of the school environment. *American Journal of Community Psychology*, *56*(3-4), 280-292.

Bradshaw, C. P., Pas, E., Bloom, J., Barrett, S., Hershfeldt, P., Alexander, A., McKenna, M., Chafin, A. E., & Leaf, P. (2012). A state-wide collaboration to promote safe and

supportive schools: The PBIS Maryland Initiative. *Administration and Policy in Mental Health and Mental Health Services Research, 39*(4), 225-237.

Bradshaw, C. P., Pas, E. T., Debnam, K. J., & Lindstrom Johnson, S. (2015). A focus on implementation of Positive Behavioral Interventions and Supports (PBIS) in high schools: Associations with bullying and other indicators of school disorder. *School Psychology Review*, *44*(4), 480-498.

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

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Cohen, J., McCabe, L., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *Teachers College Record*, *111*(1), 180-213.

Cornell, D., Sheras, P., Gregory, A., & Fan, X. (2009). A retrospective study of school safety conditions in high schools using the Virginia threat assessment guidelines versus alternative approaches. *School Psychology Quarterly*, *24*(2), 119-129.

Cornell, D., Shukla, K., & Konold, T. R. (2016). Authoritative school climate and student academic engagement, grades, and aspirations in middle and high schools. *AERA Open*, *2*, 1-18.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104.

Crosnoe, R., Johnson, M. K., & Elder Jr, G. H. (2004). Intergenerational bonding in school: The behavioral and contextual correlates of student-teacher relationships. *Sociology of Education*, *77*(1), 60-81.

Embretson, S. E. & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum Associates.

Espinoza, G., & Juvonen, J. (2011). Perceptions of the school social context across the transition to middle school: Heightened sensitivity among Latino students? *Journal of Educational Psychology*, *103*(3), 749-758.

Fan, W., Williams, C. M., & Corkin, D. M. (2011). A multilevel analysis of student perceptions of school climate: The effect of social and academic risk factors. *Psychology in the Schools*, *48*(6), 632-647.

Haynes, N. M., Emmons, C., & Ben-Avie, M. (1997). School climate as a factor in student adjustment and achievement. *Journal of Educational and Psychological Consultation*; 8(3), 321-329.

Hedges, L. V. *(1981).* Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*. 6 (2): 107-128.

Henry, D. B., Farrell, A. D., Schoeny, M. E., Tolan, P. H., & Dymnicki, A. B. (2011). Influence of school-level variables on aggression and associated attitudes of middle school students. *Journal of School Psychology, 49*(5), 481-503.

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.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. New York, NY: Routledge.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008). A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology*, *100*(1), 96-104.

Kuperminc, G. P., Leadbeater, B. J., Emmons, C., & Blatt, S. J. (1997). Perceived school climate and difficulties in the social adjustment of middle school students. *Applied Developmental Science*, *1*(2), 76-88.

Lindstrom Johnson, S., Waasdorp, T., & Bradshaw, C. (in press). School climate. In *Encyclopedia of Education*. Abindgon, UK: Routledge.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713-732.

McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum Associates.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728-743.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016-1031.

Musu-Gillette, L., Zhang, A., Wang, K., Zhang, J., Kemp, J., Kiliberti, M. and Ouderker, D.A. (2018). *Indicators of school crime and safety: 2017.* Washington, DC: National Center for Education Statistics.

National Center on Safe and Supportive Learning Environments (2018). *School climate.* Retrieved from https://safesupportivelearning.ed.gov/safe-and-healthy-students/school-climate

Payne, A. (2018). *Creating and sustaining a positive and communal school climate: Contemporary research, present obstacles, and future directions.* Washington, DC: US Department of Justice.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Schneider, S. H., & Duran, L. (2010). School climate in middle schools: A cultural perspective. *Journal of Research in Character Education, 8*(2), 25-37.

Shaw, N. (2013). *States use school score cards to target climate problems.* Retrieved from https://www.edweek.org/ew/articles/2013/03/27/26climate.h32.html

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.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 159-184). Mahwah, NJ: Erlbaum Associates.

U.S. Department of Education. (2009). *Safe and Supportive Schools Model*. Available at: http://safesupportiveschools.ed.gov/index.php?id=33. Accessed March 11 2013.

Waasdorp, T. E., Pas, E. T., O'Brennan, L. M., & Bradshaw, C. P. (2011). A multilevel perspective on the climate of bullying: Discrepancies among students, school staff, and parents. *Journal of School Violence*, *10*(2), 115-132.

Way, N., Reddy, R., & Rhodes, J. (2007). Students' perceptions of school climate during the middle school years: Associations with trajectories of psychological and behavioral adjustment. *American Journal of Community Psychology*, *40*(3-4), 194-213.

Table 1.
Sample demographics

	Total Sample ***(N* = 69,513*)***
Gender	
Female	32,127 (49.7%)
Male	32,513 (50.3%)
Race	
Native American/American Indian	1,536 (2.4%)
White	31,556 (48.8%)
Hispanic/Latino	6,230 (9.6%)
Asian/Pacific Islander	3,462 (5.4%)
Black/African American	16,628 (25.7%)
Native Hawaiian/Other Pacific	385 (0.6%)
Other	4,849 (7.5%)
Grade	
6th	11,069 (17.1%)
7th	10,006 (15.5%)
8th	8,645 (13.4%)
9th	9,957 (15.4%)
10th	9,528 (14.7%)
11th	8,658 (13.4%)
12th	6,807 (10.5%)
Maternal Education	
Did not graduate high school	2,468 (8.8%)
Graduated high school	5,716 (20.4%)
Attended some college	4,274 (15.3%)
Graduated college	15,562 (55.5%)
Grades Last Report Card	
Mostly A's	27,604 (42.8%)
Mostly B's	23,489 (36.4%)
Mostly C's	10,377 (16.1%)
Mostly D's	2,201 (3.4%)
Mostly F's	886 (1.4%)

School Characteristics (*N* = 111 schools)	***M (SD)***
% Suspension	11.6 (10.5)
School Enrollment	1059.1 (429.6)
% FARMS	39.2 (18.0)

Table 2.
Descriptive statistics and IRT parameter estimates for final scale items

	Mean	SD	b_1	b_2	b_3	a
SAFETY						
I feel safe at this school	2.07	0.73	-2.88	-1.63	1.01	1.37
I feel safe going to and from this school	2.21	0.72	-3.54	-2.19	0.70	1.12
Students at this school try to stop bullying	1.42	0.91	-1.76	0.07	2.40	1.02
Seen someone else being bullied*t	0.49	0.50	0.04	NA	NA	0.96
Harassment or bullying of students problem*	1.44	1.03	-1.01	0.05	1.23	1.86
Physical fighting between students problem*	1.59	0.97	-1.54	-0.15	1.29	1.53
Students carrying guns or knives*	2.47	0.88	-2.01	-1.47	-0.56	2.01
Programs to deal with violence and conflict*	1.64	0.90	-3.17	-0.67	2.66	0.66
Students drug use problem*	1.73	1.19	-1.00	-0.26	0.43	1.82
Students alcohol use problem*	1.96	1.16	-1.33	-0.59	0.10	1.60
ENGAGEMENT						
My teachers tell me when I do a good job	1.97	0.84	-1.86	-0.93	0.75	2.16
I enjoy learning at this school	1.74	0.93	-1.34	-0.56	1.01	2.26
My teachers believe that I can do well in school	2.20	0.78	-2.11	-1.38	0.38	2.30
My teachers listen when I have something to say	1.91	0.83	-1.67	-0.81	0.85	2.63
I feel like I belong	1.87	0.90	-1.59	-0.76	0.86	2.07
Materials reflect my culture, ethnicity, and identity	1.65	0.90	-1.98	-0.46	1.62	1.24
My teachers care about me	1.95	0.83	-1.55	-0.83	0.72	3.40
Students trust one another	1.54	0.90	-1.56	-0.22	1.67	1.50
Do good at school, my parents hear about it	1.51	1.03	-1.30	-0.04	1.30	1.39
Students and staff feel pride in this school	1.87	0.89	-1.68	-0.73	0.86	2.07
ENVIRONMENT						
Students listen to the teachers	1.45	0.81	-1.41	-0.06	1.99	2.01
Teachers can handle students who disrupt the class	1.56	0.85	-1.39	-0.23	1.49	2.34
Students are rewarded for positive behavior	1.47	0.92	-1.44	-0.03	1.76	1.42
Everyone knows what the school rules are	1.84	0.85	-2.11	-0.78	1.22	1.47
Teachers at this school help students with their problems	1.85	0.85	-1.87	-0.75	1.07	1.85
The school is usually clean and well-maintained	1.53	0.90	-1.58	-0.29	1.92	1.33
It is easy for teachers at my school to control the students	1.37	0.85	-1.25	0.16	1.87	1.97
Broken windows, doors, or desks in this school*	1.94	0.90	-3.37	-1.27	1.20	0.80
There are clear rules about student behavior	1.98	0.81	-2.18	-1.11	0.94	1.65
Misbehaving students get away with it*	1.44	0.87	-2.31	0.15	2.85	0.84

Note. b= threshold values or location parameter, α= discrimination parameter

*Items were reverse coded. [t]Item was scored dichotomously (i.e., 0/1). All other items were scored 0-3.

Table 3.
Model fit and reliability statistics for each of the three scales

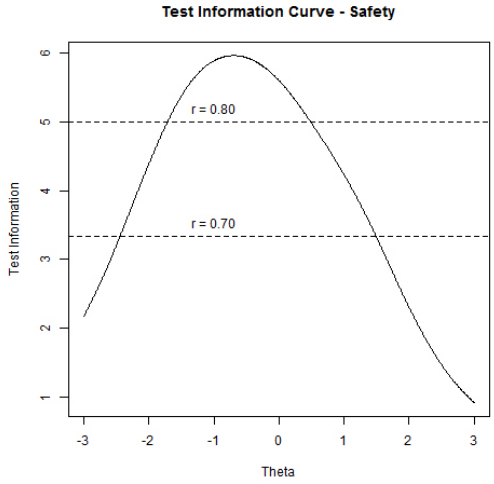
Model	CFI	SRMSR	RMSEA (CI)	M^2 (*df*)	α	ω_t
Safety	0.954	0.108	0.058 (.056-.060)	3631.94 (17)	0.80	0.86
Engagement	0.929	0.048	0.075 (.073-.076)	5063.65 (15)	0.89	0.92
Environment	0.867	0.061	0.079 (.077-.081)	5504.27 (15)	0.82	0.86

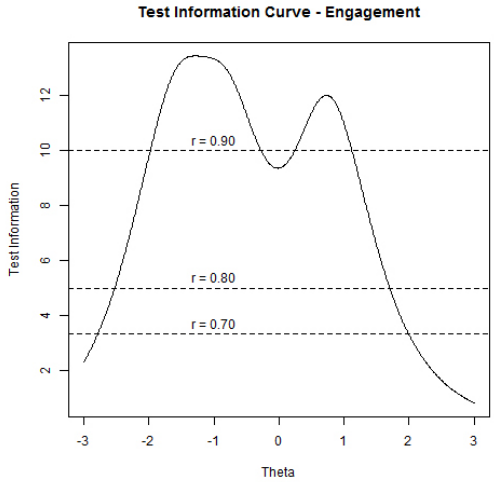
Table 4.
DIF by subgroup

DIF ESTIMATEs							
		Safety		Engagement		Environment	
Ref. Group	Focal Group	STDS	ETSSD	STDS	ETSSD	STDS	ETSSD
Male	Female	-0.156	-0.034	0.239	0.040	-0.004	-0.001
Non-White	White	-0.113	-0.024	0.097	0.017	0.206	0.043
A/B Grade	C/D/F Grade	-0.585	-0.119	-0.229	-0.037	0.054	0.011
Middle	High	-0.059	-0.013	0.083	0.014	0.014	0.003
Mat. Ed. < HS	Mat. Ed. ≥ HS	-0.101	-0.021	0.063	0.010	0.068	0.014
EFFECT SIZES OF DIFFERENCES							
		Mean Diff.	Effect Size	Mean Diff.	Effect Size	Mean Diff.	Effect Size
Male	Female	0.125	-0.138	0.082	-0.086	0.058	-0.062
Non-White	White	-0.085	0.094	-0.127	0.134	-0.138	-0.002
A/B Grade	C/D/F Grade	0.227	-0.251	0.420	-0.449	0.319	-0.347
Middle	High	-0.093	0.103	-0.023	0.024	-0.016	0.017
Mat. Ed. < HS	Mat. Ed. = HS+	-0.154	0.171	-0.163	0.173	-0.132	0.143

Note. STDS= Signed Test Difference in the Sample, ETSSD= Expected Test Score Standardized Difference, Mat. Ed= Maternal education

Figure 1.
Test information curves by scale


Test Information Curve - Safety
Test Information
Theta
r = 0.80
r = 0.70
1
2
3
4
5
6
-3
-2
-1
0
1
2
3


Test Information Curve - Engagement
Test Information
Theta
r = 0.90
r = 0.80
r = 0.70
2
4
6
8
10
12
-3
-2
-1
0
1
2
3

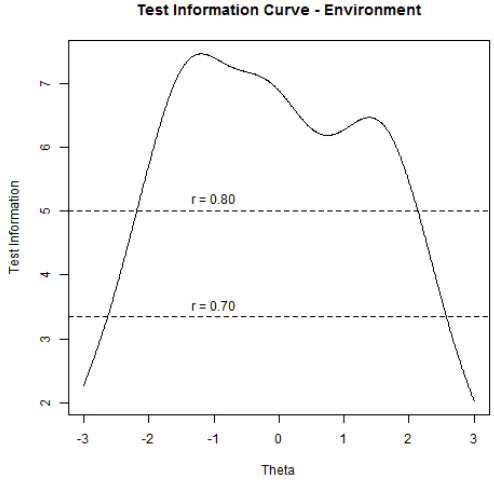
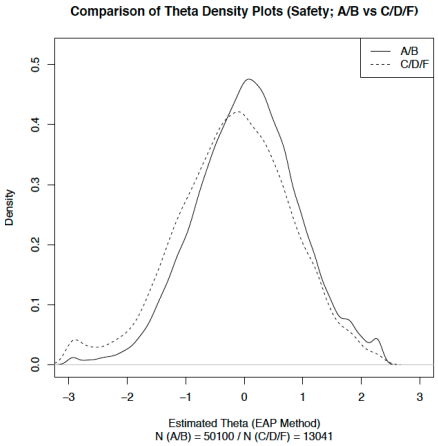

Test Information Curve - Environment
Test Information
Theta
r = 0.80
r = 0.70
2
3
4
5
6
7
-3
-2
-1
0
1
2
3

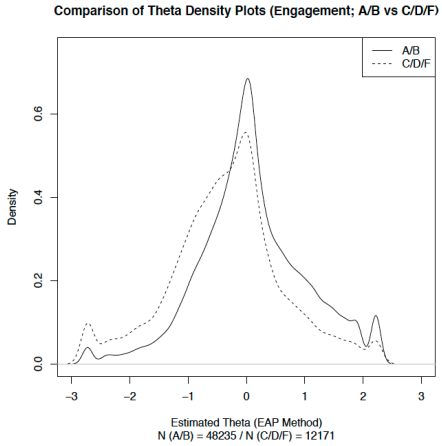
Figure 2.

Selected Comparisons of Theta Distributions by Subgroup

a) Safety



b) Engagement



c) Environment