Development of a School Survey and Index as a School Performance Measure in Maryland: A REL–MSDE Research Partnership

Tim Kautz, Charles Tilley, Christine Ross, and Natalie Larkin
June 2020

REL MID-ATLANTIC
Regional Educational Laboratory

Fueled by evidence on the strong relationships between school climate and academic achievement, teacher satisfaction, health outcomes, and social-emotional skills, states and districts are increasingly trying to measure school climate (Brand, Felner, Shim, Seitsinger, & Dumas, 2003; Gase et al., 2017; Lacireno-Paquet, Bocala, & Bailey, 2016; Voight & Hanson, 2017). School climate encompasses both tangible and intangible attributes, including relationships among students and staff, school discipline, student engagement, and safety. The Maryland State Department of Education (MSDE) partnered with Regional Educational Laboratory (REL) Mid-Atlantic to co-develop, validate, and benchmark a school climate index based upon the Maryland School Survey. The climate index will serve as a measure of school quality and student success in Maryland's school accountability framework. MSDE administered the survey statewide for accountability purposes beginning in spring 2019 following a field test in fall 2018.

This report details an approach to examining survey reliability and validity as well as converting each individual respondent's answers on the survey to an overall measure of climate for each school in Maryland. After validating the survey against standard criteria of reliability and validity, the study team used a Rasch model to develop benchmarks for each topic in the survey along four categories of school climate favorability. Based on these benchmarks, each respondent's survey responses were converted to a 1-to-10–point scale score for each topic in the survey. These topic scores were combined into an overall school climate index separately for students and instructional staff. Maryland is one of the first states to develop a measure of school climate for the state's school accountability system, and its experience may serve as a guiding example for other states and education agencies.

I. Background

In April 2017, the Maryland state legislature passed a law requiring the state's Every Student Succeeds Act (ESSA) plan to include a survey-based measure of school climate for accountability purposes. To meet this requirement, the Maryland State Department of Education (MSDE) developed the Maryland School Survey (MSS), a school climate survey for school staff and students.[1] School climate includes the social-emotional characteristics of a school community and physical characteristics of the school that can potentially influence social and academic development (Voight & Hanson, 2017; Cohen, McCabe, Michelli, & Pickeral, 2009). These characteristics include the quality of educator–student and peer relationships, expectations and support for learning, degree of connectedness and of safety that students experience in school, and the physical conditions of the school. Climate surveys can be used to collect information about the social and physical environment of a school or classroom from stakeholders, including students, parents, and staff (Voight & Hanson, 2012). Such surveys might ask respondents about their level of agreement with statements about the school or classroom, or about their own behaviors, attitudes, and perceptions. This information will allow MSDE as well as district and school stakeholders to better assess the perceptions and experiences of students and instructional staff.

Several school climate surveys have been developed over the past few years, notably surveys developed for the U.S. Department of Education (National Center for Education Statistics, 2015) and for California's CORE districts (Hough, Kalogrides, & Loeb, 2017; WestEd, 2011) as well as the Tripod/7Cs student survey, which research has shown to be predictive of student achievement growth at the classroom level (Chaplin, Gill, Thompkins, & Miller, 2014; Kane, 2012). A handful of Maryland high schools have experience with the Maryland Safe and Supportive Schools Survey (MDS3), which Johns Hopkins University and MSDE developed through a grant from the U.S. Department of Education.

[1] The Maryland School Survey (MSS) includes separate survey forms for instructional staff, elementary and middle school students, and high school students.

The MSS draws upon items from several existing surveys. The content is closely connected across two types of respondents: instructional staff and students. Separate results from these respondents will allow policymakers and educators to identify schools that have a disconnect between instructional staff perceptions and student perceptions.

The MSS provides information about the climate and quality of the learning environment of every public school in the state. This information is a statutory requirement of Maryland's accountability system as one measure of school quality and student success (SQSS), complementing other SQSS measures as well as academic indicators. A secondary purpose of the survey is to allow local educators and policymakers to shed light into the "black box" of school performance, helping to diagnose problems and design solutions to address factors that influence school performance that are otherwise difficult to observe and measure. For this reason, schools will also receive actionable information on different aspects of school climate that will inform school improvement in particular areas. The published climate measures are intended to be useful to, and used by state education agency staff, district officials, school principals, parents, and the public. Maryland's experience may serve as an example for other states and education agencies. The procedures and methodologies detailed here were informed by Maryland's goals and accountability framework, so may not be applicable in all settings.

The remaining sections of the report describe a four-step approach to developing a school climate index based on survey responses from students and instructional staff. Section II provides information on the MSS and its development. Section III describes the process of validating the survey, a newly developed instrument that had not previously been shown to be reliable and valid for students and instructional staff across Maryland. Section IV explains how the study team used a Rasch modeling approach to convert survey responses from students and instructional staff into summary measures that reflect each respondent's underlying perceptions of school climate. For each respondent, this approach yielded a score for each topic on the survey, which is then translated to a 1-to-10–point scale that aligns with Maryland's accountability framework. Section V details a process for combining these scores into a single climate index for each school. Additional methodological details are in the appendix.

II. Survey description

Both the student and instructional staff survey forms cover four broad domains with scales representing topics within each domain, as well as a separate Instructional Feedback topic for instructional staff only (Table 1). The topics and domains were selected to measure dimensions of school climate appropriate for an accountability system because they (1) reflect aspects of a school that are associated with student success, (2) represent a quality or characteristic that schools can feasibly influence, and (3) apply to a diverse set of schools. To identify these topics and domains, the study team examined the literature on school climate and convened district representatives to discuss which aspects were most relevant to their districts. After identifying these topics and domains, the study team selected corresponding items, drawing on several existing surveys, including the Delaware School Climate Survey, the ED School Climate Surveys, the Illinois 5 Essentials Survey, and a survey developed as part of a federally funded evaluation of a principal professional development program. Several surveys were used, because no single survey adequately captured all of the topics and domains identified as important. For the student school survey, each topic includes between five and nine items. For the instructional staff school survey, each topic includes between five and thirteen items. All items have the following response categories: (4) Strongly agree, (3) Agree, (2) Disagree, and (1) Strongly disagree.

Table 1. Domains and topics of the Maryland School Survey

Domain	Topic
Relationships	Student–staff relationships
	Student–student relationships
Environment	Physical environment
	Behavioral and academic supports
Community	Respect for diversity
	Participation and engagement
Safety	Physical safety
	Emotional safety
	Bullying
	Substance abuse
Instructional feedback (instructional staff only)	Quality of instructional feedback

Source: Maryland School Survey 2018-19.

Three distinct survey instruments were administered for the following populations: elementary and middle school students, high school students, and instructional staff (Table 2). The student survey forms were available in both English and Spanish. Elementary/middle and high school samples were analyzed separately for two reasons: (1) validity can differ based on the grade of the respondent and (2) the high school survey has one additional item in the Bullying topic.

Table 2. Definitions of survey respondent categories

Respondent category	Definition
Elementary/middle school students	Students in grades 5–8 who are enrolled during the survey administration
High school students	Students in grades 9–11 who are enrolled during the survey administration
Instructional staff	Staff who are classified as teachers, instructors, or other instructional personnel and were employed in the school year

Source: Authors' analysis based on Maryland State Department of Education 2018-19 data.

III. Survey validation

To meet the goal of developing a school climate measure suitable for an accountability system and school improvement, the survey must be reliable and valid. Reliability refers to the extent to which a measure consistently produces the same results if the underlying construct, such as the quality of student-teacher relationships, is the same. Validity refers to the extent to which a measure captures what it is designed to measure.

The study team followed a two-stage process to ensure that the final survey exhibited sufficient reliability and validity.[2] First, the survey was field tested statewide during fall 2018, allowing the study team to examine the properties of the survey and adjust the survey instruments to ensure sufficient reliability and validity. Second, the study team confirmed the reliability and validity of the ultimate survey by conducting similar validation exercises using data from the spring 2019 operational survey administration.

[2] Items were designed to require a reading level commensurate with the population surveyed. Additionally, the reliability and validity analyses were constructed partially with the goal of identifying items which respondents may have difficulty understanding. For example, the study team examined items for high nonresponse rates and low item response variance. Further detail on these analyses is in Appendix A.III.

Although the MSS draws on existing surveys that have already been validated to some degree, validation analyses were conducted for four key reasons. One, the MSS draws on multiple surveys that have not been validated together, so it was unclear how the scales are related. Two, some items from existing scales were removed to help minimize burden, and the abbreviated scales had not previously been validated. Three, the surveys had not been validated specifically for student and instructional staff populations in Maryland. Finally, some of the survey items were changed to better suit the context of Maryland's schools, which could alter their reliability and validity.

Analyses of the 2018 field test data revealed that the initial version of the survey performed well but prompted some minor changes to the instruments. Using the field test data, the study team examined individual-level nonresponse, item-level response patterns, measures of reliability, and the results from a confirmatory factor analysis. The results suggested the need for a few minor changes to the survey instruments (Table 3). The spring 2019 survey incorporated these changes.

Table 3. Survey modifications and rationale based on fall 2018 field test

Modification	Sample	Rationale
Created a single topic by combining the items in two topics in the Environment domain	All samples	The two topics were highly correlated and conceptually similar.
Removed one item from the Substance Abuse topic in the Safety domain	Instructional staff	This item was not strongly associated with the other items in the same topic. Subsequent benchmarking analyses suggested that including the item would degrade the measurement system.
Rephrased various items to more closely align with other items in the same topic	All samples	Several items that were not strongly correlated with other items in the same topic contained phrasing that was not closely aligned conceptually to the topic.

Source: Authors' analysis based on Maryland State Department of Education 2018-19 data.

To assess the performance of the final survey, several measures of reliability and validity were compared to rule-of-thumb benchmarks used in the literature (Table 4). Further detail on these analyses is in Appendix A.III. By standard criteria, the final spring 2019 survey performed as reliable and valid.

Table 4. Criteria for determining reliability and validity and performance of the final survey administered in spring 2019

Domain and definition[a]	Statistic	Criterion for acceptable fit	Performance of the final survey
Item response rate	The percentage of respondents who skip each item	Each item must be skipped by fewer than 10 percent of respondents (National Center for Education Statistics, 2015)[b]	All items met this criterion.
Item response variance	The percentage of responses in each response category	Fewer than 90 percent of responses fall within a particular response category (National Center for Education Statistics, 2015)[b]	All items met this criterion.
Reliability	Cronbach's alpha (Cronbach, 1951)	≥ 0.65 to 0.70 (Bland & Altman, 2007; Taber, 2018)	All topics met the lower-bound criterion for acceptable fit (0.65). All but one met the upper-bound criterion (0.70).
Overall CFA model fit	Root mean square error of approximation (Steiger & Lind, 1980)	≤ 0.05 for a close fit ≤ 0.08 for a reasonable fit (Browne & Cudeck, 1992)	All three survey forms met the criterion for a reasonable fit.
Overall CFA model fit	Comparative fit index (Bentler, 1990)	≥ 0.90 (Brown, 2015)	Two of three survey forms met this threshold. One was acceptably close (0.88).
Overall CFA model fit	Tucker-Lewis index (Tucker & Lewis, 1973)	≥ 0.90 (Brown, 2015)	Two of three survey forms met this threshold. One was acceptably close (0.89).
Relationship between items and associated topics	Standardized factor loading for each item	≥ 0.40 for each item (Stevens, 2012)	All items met this criterion.

CFA is confirmatory factor analysis.
[a]See Appendix A.III for descriptions of each of these criteria.
[b]These criterion are based on reporting standards used by the the ED School Climate Surveys (National Center for Education Statistics, 2015).
Source: Authors' analyses based on Maryland State Department of Education 2018-19 data.

IV. Survey benchmarking

Interpretability is a key aspect of a useful and actionable accountability framework. Maryland had five main priorities for developing an interpretable index for use in accountability: (1) scores for each topic should be reported on an easy-to-interpret range (for example, a 1-to-10–point scale) and in a comparable way so that schools can prioritize topics when considering school improvement; (2) the overall index should be measured on a numerical scale, so that it can be included in the accountability framework; (3) the index should have the same numerical scale as the individual topics, so that it is clear how the individual topic scores enter the overall index; (4) the topic scores and index should be comparable across different schools and respondent types; and (5) the measures should be designed to provide valid information about topics at the school level, rather than information about individual respondents or aspects of school climate captured by individual survey items.

To create an interpretable school climate measure in support of these priorities, the study team used Rasch modeling to convert the raw survey responses into measures that could be compared to meaningful benchmarks. Under a Rasch modeling approach, each respondent is assigned a "Rasch score" that reflects the respondent's

perception of school climate about a particular topic. The Rasch scoring approach offers advantages relative to the standard, Likert-based approach of forming the average score across the items in a given topic (for example, by assigning a score of 1 to 4 for responses ranging from "Strongly disagree" to "Strongly agree," then calculating the average). In particular, the Rasch approach better accounts for response patterns and missing data within each topic, as discussed below (Dogan, 2018).

To develop the benchmarks, the study team followed the approach used to benchmark the ED School Climate Surveys (National Center for Education Statistics, 2017). The benchmarking analyses followed three steps that build on each other. First, a Rasch model was estimated for each survey and topic. Second, the estimates of the Rasch model were used to determine interpretable cut points that represent different levels of school climate perception. Third, these benchmarks were used to translate scores for each topic into a 1-to-10–point scale. These procedures resulted in interpretable scales that reflect perceptions of school climate and can enter directly into the calculation of the school climate index.

Step 1. Estimating the Rasch model

The study team estimated a Rasch partial credit model[3] separately for each survey and topic (Masters, 1982). For each item on a survey, this method models the probability of selecting a given response as a function of the respondent's perception of the school's climate.[4] For example, the more favorable the perception of school climate, the more likely a respondent is to select "Strongly agree"[5] to a particular item. Every respondent is assigned a Rasch score specific to each topic. A lower Rasch score indicates a less favorable impression of school climate, whereas a higher Rasch score indicates a more favorable impression.

The Rasch scoring approach offers advantages relative to other methods by more fully accounting for response patterns and missing data within each topic in several ways (Dogan, 2018):

- *Accounting for distance between response options.* A standard, Likert-based scoring approach implicitly assumes that the "distances" between response options are equal. For example, in such a scale, a response of "Strongly disagree" might be assigned a value of 1; "Disagree" assigned a value of 2; "Agree" assigned a value of 3; and "Strongly agree" assigned a value of 4. This scoring rule implicitly assumes that the difference between "Strongly disagree" and "Disagree" is the same as the difference between "Disagree" and "Agree." In contrast, a Rasch model does not make this assumption. Instead, the Rasch model explicitly estimates the distance between response options, thereby providing more accurate scores for respondents.
- *Accounting for item difficulty.* Rasch models also allow for different items to have different "difficulties." This terminology originates in testing literature where some items might be easier to answer correctly than others. In the context of a school climate survey, this means that the same response (for example, "Agree") on different items can hold different meaning for a respondent's perception of school climate. For example, consider two items in the same topic for which more agreement indicates a more favorable perception of school climate. The first item is more difficult if—for a given perception about that topic—students are less likely to agree with the first item than the second item. Rasch scores explicitly account for the difficulty of each item, whereas standard Likert-based scoring approaches do not.

[3] The partial credit model was selected because it allows for types of fit that are more flexible than in other Rasch models applicable in this setting, such as the rating scale model (Andrich, 1978).

[4] See Dogan (2018) for additional information on the approach used in this study.

[5] Discussion about item responses and their connection to the benchmark levels assume all items are positively valenced. A positively valenced item is one for which more agreement indicates a more favorable view of school climate. For analysis purposes, all items in the MSS were re-coded such that more favorable responses corresponded to larger numeric values.

- *Accounting for item nonresponse.* Rasch scores are constructed using only observed data and do not require a complete set of responses for each respondent. Missing responses are *not* treated as "incorrect" (for example, missing responses are not assigned the lowest possible value). Instead, respondents with missing item responses are assigned Rasch scores which account for the specific items the respondent omitted. Accommodating respondents with missing item responses is preferred to treating missing responses as "incorrect" because the model estimates are less prone to bias (De Ayala, Plake, & Impara, 2001; Lord, 1974). Additionally, Rasch modeling estimates that account for respondents with omitted items are robust to various types of missing data patterns (Waterbury, 2019). Rasch scores account for the relative difficulties of individual items even in the presence of item nonresponse.

To assess the performance of the Rasch model, the study team examined the *infit* and *outfit* mean square value (Wright, 1984; Wright & Masters, 1981). The infit and outfit statistics indicate whether the observed answers fit the model as expected. These fit statistics from the spring 2019 administration met conventional criteria (Linacre, 2002). Additional information on this analysis is in Appendix A.IV.

Step 2. Creating the benchmarks

For each topic, the study team formed benchmarks based on the estimates from the Rasch model.[6] To do so, the team estimated cut points in Rasch score units that correspond to the probability of selecting various response categories for the items within the topic. For example, the uppermost cut point corresponds to the point at which a respondent with a Rasch score higher than that cut point is most likely to report "Strongly agree" for an item in the topic. Aligning the cut points to the response categories makes them more interpretable. This approach is similar to that used for the ED School Climate Surveys (National Center for Education Statistics, 2017). Details on the procedure used to calculate the cut points are in Appendix A.IV.

The interpretation of each benchmark level corresponds directly to the answer that the respondent is most likely to choose in response to an item in the topic (Table 5). For example, respondents with a Rasch score above the high cut point are most likely to "Strongly agree" that their school has a positive school climate with respect to the particular topic. Similarly, respondents with a Rasch score below the low cut point are most likely to "Strongly disagree" or "Disagree"[7] that their school has a positive school climate with respect to the particular topic. The three cut points create four distinct levels of school climate perception. The Rasch scores remain on a continuous scale to maintain the full variation of responses when used to form the climate index. The cut points provide a way to interpret the continuous scores in terms of responses to the survey items.

[6] Benchmarks are developed based on survey response categories, as described in Dogan (2018). An alternate approach would be to benchmark responses against average responses of a broader population of survey respodents (for example, a nationwide survey). Developing benchmarkings based on survey response categories aligned with Maryland's goals to develop a valid, reliable instrument for its statewide student and instructional staff populations.

[7] The response categories "Strongly disagree" and "Disagree" are collapsed into a single response category when constructing the benchmarking levels, because the Rasch model indicated that these responses provided similar information on respondents' perceptions of school climate. This approach is consistent with the ED School Climate Surveys (National Center for Education Statistics, 2017) and is further discussed in Appendix A.IV.

Table 5. Definition and interpretation of benchmarks

Description	Interpretation	Formulation
Level 4 (Most favorable)	Most likely to "strongly agree"	Above high cut point
Level 3 (More favorable)	Most likely to "agree" but more likely to "strongly agree" than "disagree/strongly disagree"	Between middle and high cut points
Level 2 (Less favorable)	Most likely to "agree" but more likely to "disagree/strongly disagree" than "strongly agree"	Between low and middle cut points
Level 1 (Least favorable)	Most likely to "disagree/strongly disagree"	Below low cut point

Note: Interpretation assumes a positively valenced question.
Source: Authors' analysis based on Maryland State Department of Education 2018-19 data.

Step 3. Converting the scores to a 1-to-10–point scale

Because Rasch scores can be negative or positive, it was necessary to rescale the scores to fall into a more interpretable range to align with Maryland's accountability framework. Individual respondents received a topic-specific Rasch score that reflects their perception of school climate where a higher Rasch score indicates a more positive impression. To place the Rasch scores for each topic on a 1-to-10–point scale, the study team conducted a conversion using the cut points as anchors. In other words, the cut points have the same value across each of the topics. The low cut point was anchored to a 2, the middle cut point to a 5.5, and the high cut point to a 9. This rescaling means that, for example, scale scores between a 5.5 and 9 will fall into the "More favorable" category, as defined in Table 5. This category is defined the same way for each topic, allowing levels to be comparable across topics.

Using the same approach as the benchmarking analysis in the ED School Climate Surveys, the study team solved a system of linear equations so that the Rasch scores for each respondent could be translated for inclusion in Maryland's accountability framework. More information on this conversion is in Appendix A.IV. This procedure translates the topic-level Rasch scores of individual survey respondents using the cut points defined above. To maintain the full range of variation in responses, the scaled topic scores of individual respondents are not restricted to the 1-to-10 range. For example, an individual respondent could have a scaled topic score of -2 or 11.5.[8] This approach is consistent with the underlying statistical theory and is a standard feature of such conversions (National Center for Education Statistics, 2017). Ultimately, school-level topic scale scores are restricted to the 1-to-10 range for alignment with the accountability framework. Therefore, it is possible for a school to have a final topic scale score of exactly 1 even if not all respondents have a topic scale score of exactly 1. For example, if some of the respondents have a topic scale score less than 1, then the school's score could also be less than 1 before the restriction and exactly 1 after the restriction. The restricting of school-level scale scores to the 1-to-10 range is discussed in greater detail in the next section.

V. Creating the school climate index

The process for creating the school climate index involves converting the topic-specific, individual-level scores into a single index value at the school level, separately for students and instructional staff. Constructing a single climate index value offers a potentially helpful tool for stakeholders to succinctly gauge the views of students and instructional staff in a school and is required for Maryland's accountability framework. Several of the steps in the following five-step process are designed to meet Maryland's accountability reporting standards and might not apply in all situations.

[8] In the 2018-19 administration, the range of individual-level scaled scores varied across topics. For example, in the elementary and middle school student survey the individual-level scaled topic scores for the student–staff relationships topic ranged from -6.5 to 13.4, while the individual-level topic scores ranged from -14.4 to 18.1 for the bullying topic.

1. *Identify valid respondents for each topic.* The study team determined whether each respondent had a valid set of responses for each topic. To be included in the index for that topic, a respondent must have answered 50 percent or more of the items in a topic and answered at least three items within a topic. The team adapted this rule from the reporting standards used by the ED School Climate Surveys (National Center for Education Statistics, 2017). Students and instructional staff missing more items are not included in the index for that topic at all. For example, they are not given the lowest score possible.
2. *Calculate a given respondent's score for the topic.* For each respondent who has a valid score for a topic, the study team calculated a scale score on the topic. To do so, the steps outlined in Section IV were applied. Let the scale score for person i in school j of respondent type r on topic t be given by SS_{ijrt}. Respondent types can either be s for student or e for educator.
3. *Calculate school-level scores on each topic for each respondent type.* The study team then averaged across individuals within a respondent type and topic for a given school to calculate:

   $$\overline{SS}_{jrt} = \sum_{i=1}^{N_{jrt}} \frac{1}{N_{jrt}} SS_{ijrt},$$

   where N_{jrt} is the number of respondents, which could differ across topics for a given respondent type and school.
4. *Restrict topic scores to a 1-to-10–point range.* School-level scores are restricted to the 1-to-10–point range for ease of interpretation and alignment with Maryland's accountability framework. School-level topic scores are bounded at 1 on the lower extreme and bounded at 10 on the upper extreme. If a school has a score of 10 for a particular topic, this indicates that respondents are most likely to "Strongly agree" that their school has a positive climate with respect to the topic. Similarly, for a school score of 1, respondents are most likely to "Strongly disagree" or "Disagree." A school bounded score of 1 or 10 does not signify that every respondent expressed an identical, very negative or very positive impression of school climate, respectively. Rather, each school-level topic score is an overall reflection of respondents' perceptions where the benchmark levels maintain the same interpretation across topics.[9]
5. *Calculate the school-level index for each respondent type.* The study team then averaged across each of the topics to create the school-level index for each respondent type:

   $$\overline{SS}_{jr} = \sum_{t=1}^{T_r} \frac{1}{T_r} \overline{SS}_{jrt},$$

   where T_r is the number of topics for each respondent type.[10] This approach weights each topic equally, which aligned with Maryland's reporting requirements. In other applications, other weighting approaches could be used (for example, equal weights for each domain). The student and instructional staff school survey results are reported separately. The final overall scores—as well as scores by student demographic characteristics—will be publicly available through MSDE's Maryland School Report Card website.

[9] The procedure of bounding scores at 1 and 10 ensures comparability across topics and does not impact the benchmark level of the school for each topic. For example, a school in the "Most favorable" region would fall in this region irrespective of the bounding of the topic scores at the upper limit. Another byproduct of the bounding is that a school's overall school climate score (student or instructional staff) will not be overly influenced by a single topic score.

[10] As a sensitivity analysis, the study team also formed these averages using weights that account for nonresponse. The results were similar with these weights, likely due to the high response rates to the survey. Nonresponse weights were not used when creating the school-level index.

VI. Looking to the future

There is a growing body of evidence on the importance of school climate for a range of student- and staff-related measures: academic achievement, teacher satisfaction, health outcomes, and social-emotional skills, amongst others (Brand, Felner, Shim, Seitsinger, & Dumas, 2003; Gase et al., 2017; Lacireno-Paquet, Bocala, & Bailey, 2016; Voight & Hanson, 2017). Maryland is one of the first states to develop a school climate index for accountability, and its experience might serve as a helpful example for state education agencies looking to incorporate survey-based measures into their statewide accountability plans under ESSA. Other states might also consider using the process of validating a statewide survey, benchmarking responses, and constructing a school climate index to inform school improvement outside of the state's formal accountability system. Maryland plans to continue administering the MSS in future school years and will replicate these procedures to promote comparisons of the school index measures over time. Maryland has demonstrated the potential to administer a psychometrically validated survey to students and instructional staff statewide as the foundation for robust measures of school climate.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. Retrieved from https://eric.ed.gov/?id=EJ204014

Bentler, P. M. (1990). Fit indexes, lagrange multipliers, constraint changes and incomplete data in structural models. *Multivariate Behavioral Research*, *25*(2), 163–172. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26794478

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *BMJ (Clinical Research Ed.), 314*(7080), 572. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9055718

Brand, S., Felner, R., Shim, M., Seitsinger, A., & Dumas, T. (2003). Middle school improvement and reform: Development and validation of a school-level assessment of climate, cultural pluralism, and school safety. *Journal of Educational Psychology*, *95*(3), 570–588. Retrieved from https://eric.ed.gov/?id=EJ674347

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Publications.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods Research, 21*(2), 230–258. Retrieved from https://journals.sagepub.com/doi/10.1177/0049124192021002005

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools* (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from https://eric.ed.gov/?id=ED545232

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. Retrieved from https://eric.ed.gov/?id=EJ826002

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16,* 297–334. Retrieved from https://link.springer.com/article/10.1007/BF02310555

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 213–234. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3984.2001.tb01124.x

Dogan, E. (2018). An application of the partial credit IRT model in identifying benchmarks for polytomous rating scale instruments. *Practical Assessment, Research & Evaluation*, *23*(7). Retrieved from https://eric.ed.gov/?id=EJ1180135

Gase, L. N., Gomez, L. M., Kuo, T., Glenn, B. A., Inkelas, M., & Ponce, N. A. (2017). Relationships among student, staff, and administrative measures of school climate and student health and academic outcomes. *Journal of School Health*, *87*(5), 319–328. Retrieved from https://eric.ed.gov/?id=EJ1136399

Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement*. Stanford, CA: Policy Analysis for California Education. Retrieved from https://eric.ed.gov/?id=ED574847

Kane, T. (2012). Capturing the dimensions of effective teaching. *Education Next*, *12*(4). Retrieved from https://eric.ed.gov/?id=EJ994599

Lacireno-Paquet, N., Bocala, C., & Bailey, J. (2016). *Relationship between school professional climate and teachers' satisfaction with the evaluation process* (REL 2016–133). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from https://eric.ed.gov/?id=ED565674

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. Retrieved from https://www.rasch.org/rmt/rmt162f.htm

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247–264. Retrieved from https://eric.ed.gov/?id=EJ108442

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. Retrieved from https://eric.ed.gov/?id=EJ270530

National Center for Education Statistics. (2015). *ED School Climate Surveys (EDSCLS): National Benchmark Study 2016*. *Appendix D. EDSCLS Pilot Test 2015 Report*. Washington, DC: Author. Retrieved from https://eric.ed.gov/?id=ED577461

National Center for Education Statistics. (2017). *ED School Climate Surveys (EDSCLS) Psychometric Benchmarking Technical Report*. Washington, DC: Author. Retrieved from https://safesupportivelearning.ed.gov/sites/default/files/SCIRP/EDSCLS_Psychometric_Benchmarking_Technical_Report_2018-04-25.pdf

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual spring meeting of the Psychometric Society, Iowa City, IA.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. New York, NY: Routledge.

Taber, K. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. Retrieved from https://eric.ed.gov/?id=EJ1200866

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. Retrieved from https://eric.ed.gov/?id=EJ075116

Voight, A., & Hanson, T. (2012). *Summary of existing school climate instruments for middle school*. San Francisco, CA: REL West at WestEd. Retrieved from https://eric.ed.gov/?id=ED566402

Voight, A., & Hanson, T. (2017). *How are middle school climate and academic performance related across schools and over time?* (REL 2017–212). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from https://eric.ed.gov/?id=ED572366

Waterbury, G. (2019). Missing data and the Rasch model: The effects of missing data mechanisms on item parameter estimation. *Journal of Applied Measurement*, *20*(2), 1–12. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/31120433

WestEd. (2011). *California School Climate Survey, statewide results, 2008-10. What teachers and other staff tell us about our schools*. Los Alamitos, CA: WestEd Health and Human Development Program for the California Department of Education.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, *3*(1), 281–288. Retrieved from https://www.rasch.org/memo41.htm

Wright, B. D., & Masters, G. N. (1981). *The measurement of knowledge and attitude* (Research Memorandum No. 30). Chicago, IL: MESA Psychometric Laboratory, Department of Education, University of Chicago.

Appendix

A.I. Background

There are no references to the appendix in the corresponding section of the report.

A.II. Survey description

There are no references to the appendix in the corresponding section of the report.

A.III. Survey validation

Item-level response patterns

Patterns in how individuals respond to items can suggest whether individual items are performing well. The study team considered two types of response patterns:

1. *Item response variance.* Item response variance refers to the extent to which different respondents mark different response categories on a given item. If a large percentage of respondents mark the same response category for an item, then it suggests the item does not adequately distinguish between different respondents.
2. *Item response rates.* If many respondents skip a particular item, then it suggests that the item might not be performing well. For example, respondents might skip an item that is poorly phrased or difficult to understand.

For each of these types of response patterns, the study team examined the items to determine whether they met standard criteria:

1. *Item response variance*. Following the validation analyses of the ED School Climate Surveys (EDSCLS), items were considered to have insufficient response variance if more than 90 percent of responses fell in a single response category.
2. *Item response rates*. Following the validation analyses of the EDSCLS, nonresponse rates were considered to be excessive if they exceeded 10 percent.

Based on the final version of the survey administered in spring 2019, all items exhibited adequate levels of response variance and nonresponse rates. Across all three respondent samples, the highest level of nonresponse for a single item was 6 percent, which is below the cutoff of 10 percent. Across all three samples, the greatest percentage of responses in a single category was 73 percent, which is below the cutoff of 90 percent.

Internal consistency

Internal consistency is a measure of reliability that evaluates the degree to which different items within a construct produce similar results. Cronbach's alpha was used to measure the internal consistency (Cronbach, 1951).

There is some debate in the field about how high a Cronbach's alpha must be for sufficient reliability, but most guidance suggests aiming for a value above 0.65 or 0.70 (Bland & Altman, 2007; Taber, 2018). Across the three survey instruments, nearly all of the scales from the spring 2019 administration meet the 0.70 criterion for adequate reliability, whereas one single scale (Participation and Engagement) achieved a reliability between 0.65 and 0.70 for the two student instruments (Table A1).

Table A1. Scale reliabilities for spring 2019 data

Domain	Topic	Elementary/ middle school students	High school students	Instructional staff
Relationships	Student–staff relationships	0.85	0.83	0.85
	Student–student relationships	0.85	0.86	0.90
Environment	Physical environment	0.81	0.81	0.81
	Behavioral and academic supports	0.83	0.82	0.91
Community	Respect for diversity	0.75	0.76	0.81
	Participation and engagement	0.69	0.66	0.80
Safety	Physical safety	0.75	0.72	0.86
	Emotional safety	0.84	0.82	0.83
	Bullying	0.82	0.83	0.81
	Substance abuse	0.89	0.88	0.89
Instructional feedback (instructional staff only)	Quality of instructional feedback			0.94

Source: Authors' analysis based on Maryland State Department of Education 2018-19 data.

Confirmatory factor analysis

In a confirmatory factor analysis (CFA), the researcher defines a structural equation model that imposes a set of assumptions about the relationships between items in a survey and conducts analyses to see whether the data support those assumptions. In the case of validating the scales of the Maryland School Survey (MSS), this approach required assuming that groups of items in the same topic capture the same underlying construct and items from different topics capture different underlying constructs.[11] For example, two items from the Bullying topic are assumed to capture the same construct, whereas items from the Bullying and Instructional Feedback topics are assumed to capture different constructs. A CFA more generally sheds light on construct validity—that the measures capture the intended constructs. For example, if two separate constructs are highly correlated with each other, then it suggests that they do not capture distinct constructs as intended. A CFA is appropriate when validating existing scales or those based on theory, as in the MSS.

The study team used the weighted least squares with mean and variance adjustment (robust) estimator (WLSMV), which has been shown to be both robust and feasible for models with categorical measures and relatively high numbers of factors, as in this context (Brown, 2015). The survey responses were modeled as categorical variables, rather than continuous ones. The CFA produced the following two types of results that help illuminate whether the groupings of items in topics are appropriate:

1. *Overall model fit.* Overall model fit statistics are a summary of whether a structural equation model fits the data well. In the context of a CFA, they provide evidence on whether the items generally group into the topics specified in the model assumptions. However, a poor model fit does not necessarily indicate that the groupings of items into topics are inappropriate, because the model might not fit the data well for reasons that are unrelated to the grouping of items into topics.
2. *Standardized factor loadings*. Structural equation models also provide an estimate of the factor loading for each item—a measure of the extent to which the item relates to the underlying factor corresponding to each

[11] The study team also allowed correlations between the error terms for a few pairs of items within the same topics based on similarities in the items' phrasing.

topic. Standardized factor loadings typically range from –1 to 1. A factor loading close to zero indicates that an item might not belong with a particular topic. Removing items with near-zero factor loadings from a scale can improve the reliability of the scale.

The study team employed standard thresholds used in the literature to assess whether the model had acceptable fit (Table A2). If the model did not meet these criteria, the team examined whether the model or survey should be adjusted as a result. For nearly all statistics examined, the model met these standard thresholds in the spring 2019 administration. The comparative fit index and Tucker-Lewis index values for the high school student survey (0.88 and 0.89, respectively) approached but did not meet the threshold values. These criteria are more stringent than those used to validate the original EDSCLS and other surveys, so they represent a conservative approach. These criteria are also all based on rules of thumb, which can be relaxed in some cases.

Table A2. Fit statistics and criteria for acceptable fit

Domain	Statistic	Criterion for acceptable fit
Overall model fit	Root mean square error of approximation (Steiger & Lind, 1980)	≤ 0.05 for a close fit ≤ 0.08 for a reasonable fit (Browne & Cudeck, 1992)
Overall model fit	Comparative fit index (Bentler, 1990)	≥ 0.90 (Brown, 2015)
Overall model fit	Tucker-Lewis index (Tucker & Lewis, 1973)	≥ 0.90 (Brown, 2015)
Relationship between items and associated topics	Standardized factor loading for each item	≥ 0.40 for each item (Stevens, 2012)

A.IV. Survey benchmarking

Infit and outfit statistics

The study team assessed the item fit for each sample by examining the *infit* and *outfit* mean square value (Wright, 1984; Wright & Masters, 1981). These statistics indicate whether the observed responses fit the model as expected. A value of the infit and outfit close to 1 indicates that the data fit the model as expected. A value greater than 1 indicates that there is more variation in the answers than the model would have predicted, and a value lower than 1 indicates that there is less variation. For example, a value of 1.2 indicates that there is 20 percent more variation in responses than would be expected. The infit and outfit statistics were assessed using the criteria in Table A3.

Table A3. Criteria for assessing infit and outfit statistics

Range of infit or outfit mean square value	Description	Recommended action
> 2.0	Distorts or degrades measurement system	Drop item from index
1.5 to 2.0	Unproductive for measurement, but not degrading	Include item
0.5 to 1.5	Productive for measurement	Include item
< 0.5	Less productive for measurement	Include item

Source: Linacre (2002).

The ideal level of fit is close to 1 (between 0.5 and 1.5 for a rating scale). The items in the spring 2019 administration fit the Rasch model well and had fit statistics that placed them in the range that indicates that they are productive for measurement. In the fall 2018 field test, one item from the Substance Abuse scale threatened

to distort the measurement system by exceeding the threshold value of 2.0. This item was excluded from the spring 2019 analysis.

Item step values

The item step values are estimates of the point at which respondents switch from being more likely to select one response category relative to the adjacent category. Because the MSS has four possible response categories, there are three corresponding step values: (1) one that separates "Strongly disagree" and "Disagree"; (2) one that separates "Disagree" and "Agree"; and (3) one that separates "Agree" and "Strongly agree." If a respondent's Rasch score is exactly at the step value, then that respondent is equally likely to select either category. For example, if a respondent's Rasch measure is exactly at the first step value, then the respondent is equally likely to select "Strongly disagree" and "Disagree."

The estimated step values help conceptualize the information provided by respondents who select one category versus another and can inform how to set benchmarks. If two step values are close to each other, then it suggests that the corresponding response categories do not provide much additional information vis-à-vis the other. For most items on the fall 2018 field test and spring 2019 survey, the distance between the lowest ("Strongly disagree") and second lowest ("Disagree") category was relatively small. This suggested that the "Disagree" category provided little measurement information not already captured by the "Strongly disagree" category. For this reason, the bottom two categories were collapsed into a single "Strongly disagree/Disagree" category for all subsequent analysis. This approach is consistent with the ED School Climate Surveys (National Center for Education Statistics, 2017). The decision to collapse response categories was empirically based and may not improve measurement interpretability for a different survey instrument or respondent sample with different properties.

Calculating cut points

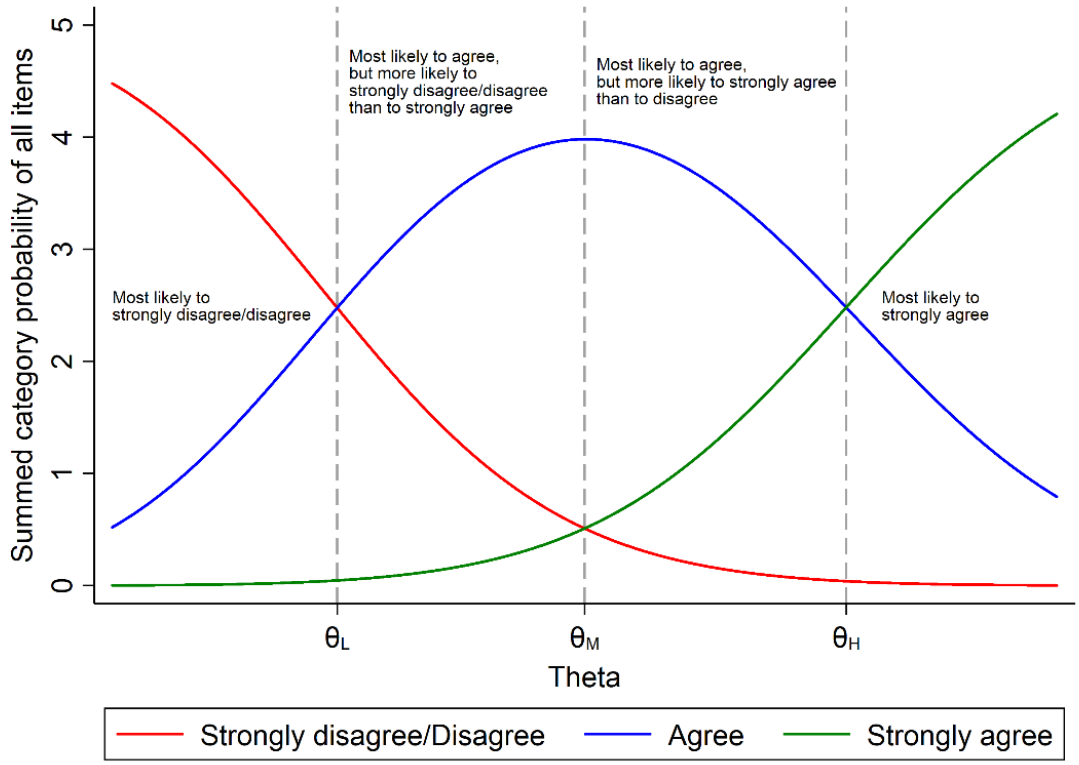
The study team used a three-step procedure to calculate the topic-specific cut points that are used to define the benchmark levels:

1. For each response category for each item, the category probability curve (CPC) was calculated. The CPC is the probability of selecting a particular category (for example, "Strongly agree") as a function of each respondent's perception of school climate θ_i.[12] For example, for a positively valenced item, a respondent is more likely to select "Strongly agree" if the respondent has a higher value of θ_i .
2. For each response category for each topic, the scale characteristic curve (SCC) was calculated. The SCC is the sum of the CPCs across items within a topic. It represents the expected number of items within a topic for which a respondent will select a given response category. An illustrative SCC for a hypothetical topic with five items is shown in Figure A1. For example, the green line in the figure displays the expected number of items for which a respondent will select "Strongly agree" for a given value of θ_i. In this case, the number is increasing as a function of θ_i. For very high levels of θ_i, Figure A1 indicates that a respondent is likely to "Strongly agree" with all five items. The cutpoints calculated in these analyses apply to this survey and population, and might differ in other applications.
3. Cut point calculations were based on the intersections between the SCCs. The intersection between two curves indicates the point at which a respondent is likely to select an equal number of responses from each category. For example, the intersection between the blue and green curves represents the value of school

[12] The study team adopted the notation of theta (θ) to represent a respondent's underlying perception of school climate (Rasch score).

climate perception (θ_H) for which a respondent is expected to select "Agree" and "Strongly agree" for an equal number of items. Let $\widehat{\theta_L}$, $\widehat{\theta_M}$, and $\widehat{\theta_H}$ be the estimates of the low, middle, and high cut points.

Figure A1. Illustrative scale characteristic curve



Source: Authors' analysis based on Maryland State Department of Education 2018-19 data.

Converting the scores to a 1-to-10–point scale

To produce an interpretable 1-to-10–point scale for schools, the study team performed a transformation of the Rasch scores using a set of linear equations. The transformation involves multiplying the respondent's Rasch score by a slope parameter (a_1 or a_2) and adding a constant (b_1 or b_2). The transformation was conducted separately depending on whether the score was above or below the middle cut point ($\hat{\theta}_M$). For example, if the original score was less than or equal to $\hat{\theta}_M$, then a_1 and b_1 were used to calculate the respondent's scale score. As discussed in Section IV, the respondents' scale scores were not restricted to a 1-to-10–point range, but the scores for schools were restricted to a 1-to-10–point range.

For each topic, the study team solved the following equations in terms of parameters that define the transformation (a_1, a_2, b_1, and b_2). This yielded values of a_1, a_2, b_1, and b_2 that can be used to perform this transformation.

$$a_1 \times \widehat{\theta_L} + b_1 = 2$$

$$a_1 \times \widehat{\theta_M} + b_1 = 5.5$$

$$a_2 \times \widehat{\theta_M} + b_2 = 5.5$$

$$a_2 \times \widehat{\theta_H} + b_2 = 9$$

Using the values of a_1, a_2, b_1, and b_2, an individual's Rasch score ($\hat{\theta}_i$) can be translated into a scale score (SS_i) using the following rules:

$$SS_i = a_1 \times \hat{\theta}_i + b_1 \; if \; \hat{\theta}_i \leq \hat{\theta}_M$$

$$SS_i = a_2 \times \hat{\theta}_i + b_2 \; if \; \hat{\theta}_i > \hat{\theta}_M$$

A.V. Creating the school climate index

There are no references to the appendix in the corresponding section of the report.

A.VI. Looking to the future

There are no references to the appendix in the corresponding section of the report.