|  |  |  |
| --- | --- | --- |
| ESE logo | | |
|  | 2018 Views of Climate and Learning (VOCAL) Validity Study | |
|  | |
| 2018 MCAS Questionnaire  October 2019 | |
| Massachusetts Department of Elementary and Secondary Education  75 Pleasant Street, Malden, MA 02148-4906  Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370  www.doe.mass.edu | |
|  | |  | |
| ESE logo  This document was prepared by the  Massachusetts Department of Elementary and Secondary Education  Jeffrey Riley  Commissioner  The Massachusetts Department of Elementary and Secondary Education, an affirmative action employer, is committed to ensuring that all of its programs and facilities are accessible to all members of the public.  We do not discriminate on the basis of age, color, disability, national origin, race, religion, sex, gender identity, or sexual orientation.  Inquiries regarding the Department’s compliance with Title IX and other civil rights laws may be directed to the  Human Resources Director, 75 Pleasant St., Malden, MA 02148-4906. Phone: 781-338-6105.  © 2019 Massachusetts Department of Elementary and Secondary Education  Permission is hereby granted to copy any or all parts of this document for non-commercial educational purposes. Please credit the “Massachusetts Department of Elementary and Secondary Education.”  This document printed on recycled paper  Massachusetts Department of Elementary and Secondary Education  75 Pleasant Street, Malden, MA 02148-4906  Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370  www.doe.mass.edu  State Seal of Massachusetts | | | |

**Table of Contents**

1. **Purpose of this report**………………………………………………………….…. 1
2. **Survey design, survey administration, and profile of respondents**…….…….... 2

2.1. School climate construct………………………………………………….. 2

2.2. Survey design principles …………………………………………………. 4

2.3. 2017 pilot item and measure development……………………………….. 5

2.4. Pilot stakeholder engagement…………………………………………...... 6

2.5. School climate construct validity improvements…………………….…… 7

2.6. New item development…………………………………………………… 8

2.7. Form building …………………………… …………………………....... 9

2.8. Form linking and anchoring process………...…………………………..... 11

2.9. Administration of forms…………………….…………………………...... 13

2.10. Profile of respondents……………………………………………………... 14

1. **Data analyses procedures**………………………………………………………… 16

3.1. Rasch methodology ……………………………………………………..... 16

**4. Validity framework** ………………………………………………………………. 16

4.1. Validity framework……………………………………………………...... 16

1. **Validity evidence for VOCAL scales and sub-scales**……………………….…... 18

5.1. Content validity……………………………………………………...……. 18

5.1.1. Overall and dimension measures……………………………..…… 19

5.1.2. Practical significance of misfitting items on school climate scores. 21

5.1.3. Practical significance of misfitting items on safety scores ……….. 24

5.1.4. Reverse-score items and misfit……………………………….…… 26

5.1.5. Content validity conclusion…………………………………..…… 27

5.2 Structural validity …………………………………………………………. 28

5.2.1. Overall dimensionality data…………….…………………………. 28

5.2.2. Residuals analyses of 76-item VOCAL data…………………….... 29

5.2.3. Residual analyses of dimension/domain data………………...…… 31

5.2.4. Sub-scale dimension/bullying correlations………………………... 32

5.2.5. Structural validity conclusion……………………………………... 33

5.3. Substantive validity……………………………………………………….. 33

5.3.1. Rating scale……………………………………………………..…. 34

5.3.2. Overall VOCAL item hierarchy………………………………….... 35

5.3.3. Engagement dimension item hierarchy…………………………..... 36

5.3.4. Safety dimension item hierarchy………………………………….. 36

5.3.5. Environment dimension item hierarchy………………………..….. 39

5.3.6. Substantive validity conclusion………………………………...…. 40

5.4. Generalizability………………………………………………………..….. 41

5.4.1. Reliability evidence……………………………………………….. 41

5.4.2. Differential item functioning (DIF) analyses…………………….... 44

5.4.3. Generalizability conclusion………………..…………………..….. 48

5.5. External validity…………………………………………………………... 48

5.5.1. Student-level responsiveness…………………………………….... 49

5.5.2. School-level responsiveness and score reporting…………………. 49 5.5.3. Concurrent validity………………………………………………... 54

5.5.4. External validity conclusion…………………………………………56

5.6. Consequential validity………………………………………………………57

5.6.1. Intended outcomes…………………………………………….......... 58

5.6.2. Unintended outcomes……………………………………………..... 59

5.6.2. Consequential validity conclusion…………………………………. 61

**6. VOCAL report conclusion**……………………………………………………….. 61

1. **References**…………………………………………………………………………. 63
2. **Appendices**……………………………………………………………………..…. 66

A. VOCAL survey specification……………………………………………… 66

B. MCAS student questionnaires (VOCAL forms) ……………………….….. 67

C. Rasch model and logit unit of measurement ……………………………… 79

D. Guide for evaluating Rasch model validity data……………………………82

E. Technical quality of VOCAL scale and dimension scales………………… 83

F. Winsteps residual analyses output ……………………………………........ 88

G. Measure order of 76-item VOCAL scale………………………..……….... 89

H. Item prompts by dimension………………………………………………... 90

I. Person Reliability of VOCAL scale, grade-level scales and sub-scales ….. 95

I. Subgroup DIF plots ………………………………………………………... 96

J. Transformation of logit scores……………………………………………... 99

1. **Purpose of this report**

This report offers reliability and validity evidence to support the use of the Views of Climate and Learning (VOCAL) school climate survey developed by the Massachusetts Department of Elementary and Secondary Education (DESE). DESE sought to develop a school climate instrument that would: (1) differentiate levels of school climate within and between schools, and (2) provide schools and districts with concrete, actionable information about school climate in order to support continuous improvement. A positive school environment is associated with healthier social and emotional well-being, reduced substance abuse, and decreased student behavioral problems in school (Thapa, Cohen, Guffey and Higgins-D’Alessandro, 2013), and is positively related to students’ academic success (Berkowitz, Moore, Astor, and Benbenishty, 2017; Hough, Kalogrides, and Loeb, 2017). This technical report provides information on the survey development process used to develop three forms (grade 5, grade 8, and grade 10) of the school climate survey. The report focuses on the reliability and validity analyses performed to justify the use and reporting of the 2018 VOCAL scores to schools and districts. It complements the validity work reported previously ([DESE 2018a](http://www.doe.mass.edu/research/vocal/2017/validity-study.docx)).

This report is intended for readers with knowledge of survey development and validation, psychometrics, and educational measurement. Familiarity with Wolfe and Smith’s (2007a, 2007b) and Messick’s (1995a) construct validity frameworks for instrument development is helpful. School climate is a psychological construct; students provide their perceptions of their school climate by responding to statements in the VOCAL survey. Evidence from six aspects of construct validity (content, structural, substantive, generalizability, external, and consequential) combine to justify the use of VOCAL scores as a measure of students’ perceptions of school climate. All six validity aspects are addressed in this study; coverage of consequential validity is relatively limited when compared to the other five aspects.

1. **Survey design, survey administration and profile of respondents**

Instrument development relied on a five-pronged strategy: (1) defining the school climate construct, (2) incorporating stakeholder feedback to support item and instrument development, (3) using Rasch theory to ideate and guide item development and validity analyses, (4) piloting the VOCAL instrument in 2017, and (5) using 2018's survey administration to pilot new items designed to improve the psychometrics and reliability of the 2017 VOCAL survey. VOCAL instrument development and validity activities are summarized in Figure 1.

* 1. **School climate construct**

DESE used the United States Department of Education’s (USED, 2019) conceptual framework for the school climate construct, with survey items designed to measure student perceptions of three dimensions of school climate: engagement, safety and environment. Each dimension is further divided into three domains/topics. The engagement dimension items measure cultural and linguistic competence, teacher/adult-on-student relationships and student-on-student relationships, and participatory engagement in class and school life. Items measuring student perceptions of safety cover: emotional safety, physical safety, and bullying/cyber-bullying. The three environment domains are: instructional environment, mental health environment, and discipline environment. Items from publicly available school climate instruments were evaluated for inclusion, with school climate research articles furnishing ideas for new item development. DESE leveraged work done during the development of its educator evaluation [student feedback surveys](http://www.doe.mass.edu/edeval/feedback/surveys.html) (SFS), with several SFS items adapted for inclusion in the school climate surveys.

Figure 1: VOCAL scale development process

Figure 1 summarizes the VOCAL instrument development and validity activities. Instrument development relied on a five-pronged strategy: (1) defining the school climate construct, (2) incorporating stakeholder feedback to support item and instrument development; (3) using Rasch theory to ideate and guide item development and validity analyses, (4) pilot testing the instrument in 2017, and (5) using 2018's survey to pilot items designed to improve the VOCAL surveys.

The conceptual framework and construct domain definitions are outlined in Table 1.

Table 1

VOCAL’s conceptual framework**1**

| **Dimension** | **Domain (label)** | **Definition** |
| --- | --- | --- |
| Engagement  (ENG) | Cultural and Linguistic Competency (CLC) | The extent students feel adults/students value diversity, manage dynamics of differences, and avoid stereotypes. |
| Engagement  (ENG) | Relationships (REL) | The extent students feel there is a social connection and respect between staff/teachers and students, and between students and their peers. |
| Engagement  (ENG) | Participation (PAR) | The extent students feel engaged intellectually, emotionally, and behaviorally in the classroom, and the extent that students or their parents are engaged in school life. |
| Safety  (SAF) | Emotional Safety (EMO) | The extent students feel a bond to the school, and the extent adults/students support the emotional needs of others. |
| Safety  (SAF) | Physical Safety (PSF) | The extent that students feel physically safe within the school environment. |
| Safety  (SAF) | Bullying/Cyber-bullying (BUL) | The extent that students report different types of bullying behaviors occurring in the school and the extent that school/staff/students try to counteract bullying. |
| Environment  (ENV) | Instructional (INS) | The extent that students feel the instructional environment is collaborative, relevant, challenging and supportive of learning. |
| Environment  (ENV) | Mental Health (MEN) | The extent that students have access to support systems that effectively support their social, emotional, and mental-health well-being. |
| (ENV) | Discipline (DIS) | The extent that discipline is fair, applied consistently and evenly, and a shared responsibility among staff, teachers, and students. |

1Based on the USED’s conceptual framework (USED, 2019)

##### **Survey design principles**

The surveys were designed with the rigor expected of cognitive tests. When developing measures in the Rasch framework, best test design (Wright & Stone, 1979) consists of:

* Items that are evenly spaced from easiest to hardest;
* Average item difficulty (usually set to zero) is centered at the mean of the target or student distribution;
* Survey items that are sufficiently dispersed to cover the target distribution;
* Items from different dimensions/domains overlap each other on the item-person continuum; and,
* Enough questions to provide the responsiveness required to differentiate performance.

These psychometric criteria were adopted and used to guide the selection of items for the school climate survey. However, stakeholder engagement and feedback, discussed in the next section, also contributed substantially to item selection.

* 1. **2017 pilot item and measure development**

DESE developed items using a hierarchical perspective. DESE first identified what behaviors, practices, or systems create the foundation for a positive school climate; students have more likelihood of responding affirmatively to these foundational items. DESE then identified behaviors, practices, or systems that represent exemplary school climates. These behaviors/practices/systems, by their nature, are more difficult to enact within schools and students are likely to have greater difficulty responding affirmatively to items designed to measure them. Once these behaviors/practices/systems were identified, items were developed or acquired from publicly available surveys (see Appendix H for acknowledgements) to measure and anchor the two ends of the school climate continuum. The next step in the item development process was to develop or obtain publicly available items to fill in the continuum. Therefore, the rating scale (always true to never true) combined with the hypothesized distribution of item difficulty is designed to stretch the item calibrations and person distributions along the school climate continuum for each dimension/domain and provide meaningful differentiation of student perceptions.

Items for the grade 5 form were simplified to ensure students could read and understand the content. For example, the item, “Adults working at this school treat all students respectfully, regardless of students’ race, culture, family background, sex, or sexual orientation” was administered in grade 8 and grade 10; the corresponding item in grade 5 was, "Adults working at this school treat all students respectfully." Items were also developed for the specific school climate context. For example, the item, “I have been teased or picked on more than once because of my real or perceived (imagined) sexual preference” was only administered on the grade 10 survey. Similarly, an item related to cyber-bullying was placed on the grade 8 form to account for the predominance of this type of bullying in middle-school grades. Once items were selected or developed, they were reviewed by diverse stakeholder groups.

* 1. **Pilot stakeholder engagement**

Stakeholder engagement activities predominantly occurred during the 2017 pilot development. The detailed description of and findings from the 2017 stakeholder engagement activities can be found [here](http://www.doe.mass.edu/research/vocal/2017/validity-study.docx). Multiple stakeholder groups (agency experts, student advisory council members, principal and teacher advisory council members, and special interest groups) met in 2016 to review pilot items for the 2017 survey. The item review process also prompted new item development. Three to four times the number of items needed for the final surveys were developed or selected, and students and other stakeholders were asked to rate them. The process was designed to ensure item representativeness (did the items measure the concept they were designed to measure?), accessibility (would students understand the item?), actionability (would schools be able to use the information?), and responsiveness (would the items measure a continuum of student perceptions that differentiate relatively strong school climates from relatively weak ones?). Stakeholders worked in groups to review, revise, and reject items.

To further ensure items placed on the grade 5 form met these inclusion criteria, cognitive interviews were undertaken with a small, but diverse group of fifth-graders. The purpose of these interviews was to elicit and probe whether students understood the item content as the item developer intended. Participants in the cognitive interviews reported that most of the items were easy to understand. The interviews, however, did result in DESE simplifying the content and readability of some items. Through a deliberative process, the items that survived the review process were placed on the three forms of the school climate survey; each grade-level form was designed to meet the best survey design criteria highlighted previously. Reliability and validity analyses for the 2017 pilot study are provided [here](http://www.doe.mass.edu/research/vocal/2017/validity-study.docx). The results from the pilot indicated some deficiencies in each of the three forms. There was a lack of items that measured students who had very positive views of their school climates. In addition, some of the domains were not clearly defined. This study highlights the improvements made in 2018 to better measure the school climate construct and provides the evidence needed to support the use of the 2018 scores.

* 1. **School climate construct validity improvements**

This section outlines the major changes made to improve the VOCAL pilot survey. Refining the construct validity of the VOCAL instrument was the primary focus of survey enhancement efforts. To improve the construct validity of the school climate construct, the following changes were made to the 2017 survey:

1. The participation topic was expanded to measure the engagement of students in the classroom, not just in school life. The definition was revised to include the measurement of students’ perceptions of their cognitive, emotional, and behavioral classroom engagement.
2. The instructional environment items were more clearly delineated from the participation engagement topic. The participation topic now measures students’ perceptions of their self-engagement (cognitive, behavioral, and emotional) within the classroom, whereas the instructional environment topic is now aimed to measure how teachers create and maintain a supportive environment that fosters student engagement.
3. Emotional safety items were more clearly defined and separated from mental health environment items. Emotional safety items target students’ self-perceptions of their emotional safety within the school, whereas the mental health environment items now assess students’ perceptions of how well their school has developed “systems” to effectively support students’ social and emotional well-being.
4. The bullying/cyber-bullying topic was expanded to better address power imbalances that result in aggressive or harassing behaviors among students. Power imbalance could, for example, reflect when groups of students tease or bully individual students or a bully is bigger in size when compared to the victim. The number of items measuring bullying was also increased between 2017 and 2018 in order to provide districts with a reliable measure of the bullying climate across their schools.
   1. **New item development**

DESE leveraged its [student feedback surveys](http://www.doe.mass.edu/edeval/feedback/surveys.html) to select or adapt items for the expanded participation topic. The items chosen for VOCAL had been previously tried out with grade-level appropriate students using cognitive labs and pilot survey administrations (the [2014 technical report](http://www.doe.mass.edu/edeval/feedback/2014SFS-TechReport.pdf) is available for those seeking more information on DESE’s student feedback surveys). For example, the items, “When I am stuck, my teachers want me to try again before they help me” and “If I finish my work early, I have an opportunity to do more challenging work” were adapted and included in the VOCAL survey to measure students’ behavioral and cognitive engagement in the classroom, respectively. Supporting students to develop persistence and differentiating instruction are important engagement practices in the classroom.

The new items developed for the bullying/cyber-bullying topic centered on measuring bullying interactions that result from power imbalances in student relationships. For example, students across all three grades responded to the following item, “In my school, groups of students tease or pick on one student.” Another example of a new bullying item that relates to power imbalance is, “Students with learning or physical difficulties are teased or picked on at my school.” Five new bullying items were added to the 2018 survey. The items retained from the 2017 pilot and the new 2018 items were distributed across three forms of the survey, one each for grades 5, 8 and 10.

* 1. **Form building**

DESE administered three parallel forms of the VOCAL survey in the spring of 2018; the number of items on each form was: 36 items for grade 5 students; 38 items for grade 8 students; and 38 items for grade 10 students.

Each survey measured the breadth of the school climate construct and included common items that were used to place all student responses onto the same scale metric; common items (items that were on all three forms) represented over a third of the total number of items on each form. The number and types of items on each form are shown in Figure 2, with a detailed “test” specification found in Appendix A.

Figure 2. Form building for VOCAL surveys

This picture depicts how the three forms of the VOCAL survey (grade 5, grade 8, and grade 10) are linked using common items. 13 items are common to all three forms, 3 to G5 and G8, and 7 are common to G8 and G10.

Common items should represent the breadth of the school climate construct and approximate the average item difficulty and variance of all 76 items (Engelhard, 2013). Three items from the engagement dimension, six items from the safety dimension, and four items from the environment dimension make up the items common to all three grades. The 76 items had an average item difficulty of 0.00 logits and a standard deviation of 0.68 logits; the average item difficulty for the 13 common items was 0.24, with a standard deviation of 0.72. In order to provide districts with a reliable bullying score, five of the six safety items were from the bullying domain. To further boost the linkage between forms, and to ameliorate the over-representation of bullying items in the common items, additional items were added to the three forms; e.g. three items (1 emotional safety, 2 instructional environment) were added to the grade 5 and grade 8 forms. To reduce positioning effects, common items were placed in the same fixed position on each of the three forms. Once the common items were placed in their item slots, the remaining unique items were randomly assigned positions on each form.

A Likert scale with four response options was used to rate students’ perceptions of school climate. Coding for all positively valenced items dictated that a response of “0” (*untrue*) indicated the lowest level of school climate, with a “3” (*always* *true*) denoting the most positive school climate. Response scoring categories “1” and “2” corresponded to *mostly untrue and mostly true*, respectively. Note, sixteen items were reverse-scored: eight bullying behavior items, five physical safety items, one emotional safety item, and one mental health environment item were reverse-scored. A higher item score, irrespective of whether the item is positively or negatively valenced, is associated with a more positive school climate. The three forms with the items ordered as they appeared to students in each grade are provided in Appendix B1 (grade 5), B2 (grade 8), and B3 (grade 10). The appendices highlight, in green, the 13 common items administered on all three forms.

* 1. **Form linking and anchoring process**

**Form linking.** Each grade form was first calibrated separately to assess the invariance of the common items. The Pearson product-moment correlation (henceforth Pearson correlation) of the 13 common item difficulties was 0.90 between grade 5 and grade 8; 0.98 between grade 8 and grade 10; and 0.86 between grade 5 and grade 10, respectively. Figure 3 illustrates the linking process. Figure 4 graphically shows the relationship between the 13 items common across the three grade forms. The 3 additional items linking the grade 5 and grade 8 forms, and the 7 additional items linking the grade 8 and grade 10 forms did not impact the magnitude of these correlations. The correlation of the grade 5 and grade 8 forms for the 16 common items was 0.89; the correlation of the grade 8 and grade 10 forms for the 20 common items was 0.98. The magnitude of these correlations justified the concurrent calibration of all 76 items on to the same scale metric.

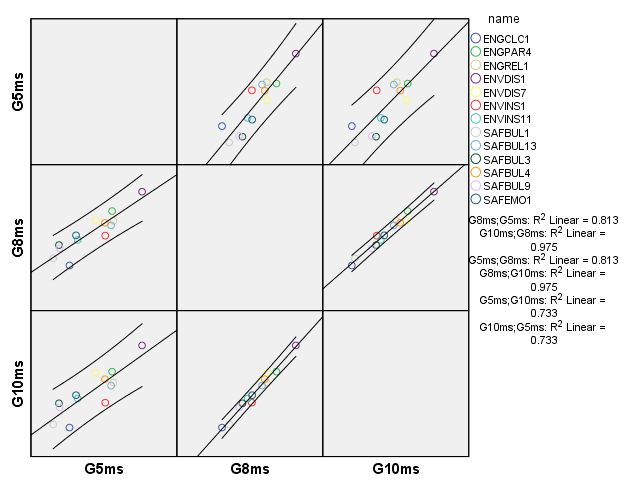
Figure 3. Concurrent calibration process of grade 5, grade 8, and grade 10 forms

Concurrent calibration process of grade 5, grade 8, and grade 10 surveys

This graphic visually portrays the concurrent calibration process used to link the three grade-level forms (grade 5, grade 8, and grade 10). The vertical dashes represent items on each form. 13 items are common to all three grade-level forms, with a further three items common to grade 5 and grade 8, and seven common to grade 8 and grade 10.

1Figure template taken from Linacre (2019); Each item is indicated by a vertical dash. To reduce positioning effects, each common item is placed in the same item position on each survey. Note: not all survey items are portrayed.

Figure 4. Relationship of 13 common items across grade forms



**Anchoring Process.** The ensuing validity analyses (and review of the item-variable map for the relative difficulty, ordering, and spacing of items), revealed that 56 of the 76 items were well-fitting and could be anchored (outfit mean square errors ranged from 0.67 to 1.40 and point-to-measures ranged from 0.31 to 0.63). To anchor the scale, the items’ average difficulty parameters and the rating scale’s Andrich step threshold parameters were both fixed (Linacre, 2019). Fifty-six items of the VOCAL scale and the rating scale structure were anchored, with the remaining 20 items allowed to float. To assess the impact of anchoring on all the items, the displacement of the items was examined. Displacement shows the difference between the item difficulty estimate when it is anchored versus when it is calibrated freely; large displacement values suggest that the anchoring process has distorted the measurement process and could lead to biased person estimates. Appendix E1 shows the displacement of all VOCAL items. Anchoring had little to no impact on the items estimates of the items that could float.

* 1. **Administration of forms**

In grades 5 and 8, the forms were administered as part of the Massachusetts Comprehensive Administration System (MCAS) Science and Technology/Engineering (STE) achievement test. In grade 10, the form was attached to the mathematics MCAS test. The forms were attached as their own test sessions on the STE or mathematics MCAS assessment. The MCAS test is administered annually to students within the three grades; schools are responsible for the MCAS and survey administration. The forms in grade 5 and grade 8 were computer-based; the form in grade 10 was paper-based. The computer-based surveys were designed to provide one item per screen; students provided their response, and then advanced to the next screen and item. Each item/screen shot was prefaced with the words, “Think of the last 30 days of school.” Grade 10 students received a paper version of the survey and marked their responses in their MCAS student answer booklets. More details of the survey administration protocols can be found [here](http://www.doe.mass.edu/mcas/testadmin/manual/PAM-g3-8.pdf).

* 1. **Profile of respondents**

The sampling frame included students in grades 5, 8 and 10. Students who participated in MCAS-Alternative were not included in the sampling frame, so a census was not attained. In addition, participation in the survey was optional for districts, schools and students. Response data indicated 82% of fifth graders, 87% of eighth graders, and 64% of tenth graders participated in the surveys, respectively. However, 3,346 grade 10 students’ responses were removed from the dataset; these students had marked their survey, but only responded to 1 or 2 random items. This reduced the grade 10 response rate to 59.5%. Of the usable surveys, over 90% of grade 5 (99.6%), grade 8 (97.2%), and grade 10 (93.3%) students fully completed their surveys. Except for grade 10 students where survey responses were not likely missing at random, no surveys were excluded due to non-response of items. The Rasch model is robust to missing data and will estimate parameters and scores based on all non-missing data available. Scores for students with a relatively high number of item data missing will have larger standard errors and, as a result, could potentially negatively impact the reliability of school-level scores. Schools only received VOCAL scores if their aggregate scores met the minimum person reliability requirement of 0.7.

The profile of the sample is reasonably representative of the state for grade 5 and grade 8, less so for grade 10; the data are shown in Table 2.

Table 2.

Participating students’ profile

| **Subgroup**  **(percent1)** | **Grade 5**  **sample** | **Grade 5**  **state** | **Grade 8**  **sample** | **Grade 8**  **state** | **Grade 10**  **sample** | **Grade 10**  **state** |
| --- | --- | --- | --- | --- | --- | --- |
| Number of students | 59,216 | 72,487 | 62,857 | 71,968 | 46,660 | 72,378 |
| Percent response | 82% | 100% | 87% | 100% | 64% | 100% |
| Usable surveys | 59,216 | 72,487 | 62,857 | 71,968 | 43,514 | 72,378 |
| Percent usable | 82% | 100% | 87% | 100% | 60% | 100% |
| Completed surveys | 58,969 | 72,487 | 64,091 | 71,968 | 40,594 | 72,378 |
| Percent complete2 | 99.6% | 100% | 97.2% | 100% | 93.3% | 100% |
| Female | 49.3 | 48.7 | 49.2 | 48.7 | 50.3 | 48.9 |
| Male | 50.7 | 51.3 | 50.8 | 51.3 | 49.6 | 51.0 |
| Non-binary3 | --- | <.01 | --- | <.01 | --- | <.01 |
| Asian | 6.6 | 6.9 | 6.3 | 6.5 | 6.1 | 6.3 |
| Black | 9.3 | 8.2 | 9.1 | 8.3 | 8.8 | 9.3 |
| Hispanic | 21.2 | 22.5 | 19.0 | 20.9 | 16.7 | 22.0 |
| Other4 | 4.0 | 4.5 | 3.5 | 4.5 | 3.1 | 3.8 |
| White | 59.0 | 57.4 | 62.1 | 59.8 | 65.3 | 58.6 |
| Students with disabilities | 17.9 | 20.1 | 16.6 | 19.0 | 14.3 | 18.4 |
| English learners | 9.1 | 8.4 | 7.4 | 6.3 | 5.6 | 8.2 |
| Economically disadvantaged | 36.7 | 36.7 | 32.8 | 32.7 | 29.1 | 33.3 |

1The number of usable surveys is the denominator; 2Percent of students who provided a response to all items on the survey; 4Includes Multi-race, Non-Hispanic, Native American, and Native Hawaiian, Pacific Islander students

Students with disabilities are under-represented in grade 5 and grade 8; Hispanic, economically disadvantaged, English learner, and student with disabilities students are all under-represented in grade 10.

1. **Data analyses procedures**
   1. **Rasch methodology**

Analyses using the Rasch measurement model (Rasch, 1960) and validity framework (Wolfe & Smith, 2007a, 2007b) are the primary source of reliability and validity data for the VOCAL survey measures. The Rasch model, which uses an exponential transformation to place ordinal Likert responses on to an equal-interval logit scale, was used to analyze student responses. Winsteps software developed by Linacre (2019) was used to perform Rating Scale model analyses of the data (Andrich, 1978a, 1978b). Technical details explaining the Rasch model are provided in Appendix C1 and C2. In the Rasch framework, the scale metric axis represents the desirable structural properties of a Rasch scale; it is: linear, unidimensional (measures only one construct), hierarchical (items are ordered according to their difficulty to affirm), and measures a continuum of items and persons. The evaluation criteria used to perform a Rasch-based reliability and validity assessment for each construct validity aspect (content, structural, substantive, generalizability, external, and consequential) are summarized in the next section.

1. **Validity Framework and Validity Evidence**
   1. **Validity framework**

Messick’s (1980, 1995a) unified concept of construct validity guided the validity analyses for the school climate construct. Messick (1995a, p. 741) defines validity as “an evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment.” Evidence from six aspects of test validity (content, structural, substantive, generalizability, external and consequential) combine to provide survey developers with the justification to claim that the meaning or interpretability of the survey scores is trustworthy for the survey’s intended use. More recently, Wolfe and Smith (2007a, 2007b, p. 205) used Messick’s validity conceptualization to detail instrument development activities and evidence that are needed to support the use of scores from instruments based on the Rasch measurement framework. Table 3 outlines the specific validity aspects addressed in this technical report.

Table 3

Rasch-Based instrument validity framework and evidence collected for VOCAL survey1

| **Validity aspect evidence** | **Validity aspect evidence** | **Validity aspect evidence** |
| --- | --- | --- |
| **Content** | **Structural** | **Substantive** |
| Instrument purpose  Test specification  Expert reviews and student focus groups/cognitive labs2  Item technical quality | Principal components residual analyses  Rasch dimensionality analyses | Rating scale functioning  Item difficulty hierarchy |
| **Validity aspect evidence** | **Validity aspect evidence** | **Validity aspect evidence** |
| **Generalizability** | **External** | **Consequential3** |
| Differential item functioning  Person separation reliability  Item invariance | Responsiveness  Sub-scale correlations  Relationship between VOCAL scaled-scores with scores from similar/dissimilar constructs | *Standard setting*  Score use |

1 Validity framework is based on: Messick (1995a) and Wolfe and Smith (2007b) conceptualization and representation.2Expert interviews, focus groups and cognitive labs were mostly carried out during the pilot phase of the survey development in 2017; 3Standard setting is not part of this study.

This report primarily focuses on internal validity with more limited external validity evidence provided for the school climate construct. Section 5 elaborates on each aspect of construct validity outlined in Table 3 and provides the evidence used to justify the use of VOCAL scores to measure school climate.

1. **Validity evidence for VOCAL scale and sub-scales**

The six aspects of construct validity (content, structural, substantive, generalizability, external, and consequential) are discussed in turn. The goal of these analyses was to ensure that DESE could report four reliable and construct-relevant scores to schools and districts (an overall school climate VOCAL score, an engagement score, a safety score, and an environment score), and one additional bullying climate topic/domain score to districts. Appendix D provides a guide to the validity criteria used in this study for each aspect of construct validity.

* 1. **Content validity**

Content validity examines the “content relevance, representativeness and technical quality” (Messick, 1995a, p.745) of the items used as indicators of the construct. Stakeholder engagement activities (Figure 1) ensured that the items were relevant and representative and, more importantly, had the potential to provide schools with diagnostic and actionable information. The content validity evidence reported here predominantly focuses on the technical quality of the VOCAL survey items. Item technical quality was assessed using point-to-measure (PTM) correlations and item fit statistics (outfit mean square error). The PTM correlations and item fit statistics are shown in Appendix E1 through E6.

PTMs below 0.3 indicate that the item is not likely construct relevant (Appendix D). The outfit mean square error fit statistic was used in this study to assess item technical quality; this statistic provides the most stable fit statistic and is least impacted by large sample sizes (Smith, 2008). Item outfit mean square error fit statistics of between 0.5 and 1.5 are productive for measurement (Wright and Linacre, 1994; Boone, Staver, and Yale, 2014; Linacre, 2019). Items whose mean square outfit statistics range between 1.5 and 2.0 have additional sources of variance but do not degrade measurement (Appendix D). Fit statistics above 2.0 are likely to degrade measurement (Wright and Linacre, 1994; Boone, Staver, and Yale, 2014; Linacre, 2019). The results from the content validity analyses are discussed next.

* + 1. **Overall and dimension measures.** When all 76 VOCAL items were calibrated together fourteen items of the seventy-six items had outfit mean square errors (MNSQ) of greater than 1.5. The results are shown in Appendix E1 (misfitting items are shown in orange). Only three of the fourteen misfitting items, however, had PTMs below 0.3 which suggests the remaining 11 items are largely related to the school climate construct. In terms of content, ten of the fourteen misfitting items were from the safety dimension and all ten required reverse scoring. Three misfitting items were from the engagement dimension, with one misfitting item from the environment dimension. The environment item was also a reverse-scored item. Given the goal of providing dimension scores for engagement, safety, and environment to schools and districts, the outfit of the items was examined when items for each dimension were calibrated separately. The outfit statistics for the engagement, safety, and environment dimensions are found in Appendix E2, E3, and E4, respectively. Any misfitting items in these tables are shown in orange.

No items remained misfitting in the engagement dimension with PTMs ranging from 0.39 to 0.64. Five items still misfit the model for the safety dimension (misfit ranged from 1.5 to 2.0); PTMs, however, varied from 0.40 to 0.60. These items differed in terms of content; they were designed to measure bullying and physically threatening behaviors. These behaviors are essential to measuring students’ views of their overall safety within the school. One reverse-scored environment item (MEN9; outfit, 1.66) still misfit the model in the environment dimension analyses; PTMs ranged from 0.31 to 0.66.

The VOCAL survey helps the state to meet requirements included in section 370 of the [Act Relative to Bullying in Schools](https://malegislature.gov/Laws/GeneralLaws/PartI/TitleXII/Chapter71/Section37O). In addition to the dimension scores, districts will receive a bullying score made up of items related to bullying behaviors (e.g., In my school older students scare or pick on younger students), and of items related to bullying protective behaviors (e.g., If I tell a teacher or other adult that someone is being bullied, the teacher/adult will do something to help). When all bullying related items were calibrated separately, the highest outfit MNSQ error was 1.51 for item SAFBUL11 (I have been teased or picked on more than once because of my race or ethnicity) and PTMs ranged from 0.53 to 0.71 (Appendix E5).

Across dimensions, the fit of these misfitting items improves when these items are calibrated separately, signifying they appropriately measure the dimension constructs and bullying topic construct. The fact that the items fit when the dimension items are calibrated separately suggests that the misfitting items are needed to productively measure the different aspects of school climate. The removal of these misfitting items from the survey could threaten the content validity and reliability of measuring the dimension constructs (Crisan, Tendeiro, and Meijer, 2017; Carmen and Johannes, 2017). Yet, the results indicate that when all 76 items are calibrated together, these items misfit the model which raises the concern that the overall school climate scores are biased. The practical significance of removing the misfitting items on the overall school climate scores was investigated. Because many of these items were safety items, the practical consequence of retaining these items for the safety dimension calibrations was also investigated. In addition, a large proportion of the misfitting items were reverse-scored items; separate content validity analyses were conducted to assess their impact.

These sets of analyses examined if the misfitting items distorted the school climate scores of students and the aggregate scores received by schools. Specifically, the impact of removing the misfitting items was assessed by: (1) evaluating the degree to which school-level scores were biased and led to the misclassification of schools; (2) comparing the number of schools who met the minimum reliability requirement (school-level person separation reliability of 0.7 or more); and, (3) by estimating any differences in student-level subgroup scores. The results of these analyses follow.

* + 1. **Practical significance of misfitting items on overall school climate scores.** Practical significance is defined as, “an assessment of the extent to which the decisions made from test scores are robust against the misfit of the IRT model,” (Sinharay & Haberman, 2014, p. 441; Van Rijn, Sinharay, Haberman, & Johnson, 2016, p. 9); these authors suggest examining and comparing the decisions made when parameters are estimated with and without the misfitting items.

**School-level score bias of removing misfitting items.** Linacre (2010) and Crisan and Tendeiro, and Meijer (2017) recommend the following empirical analyses to determine the practical significance of removing misfitting items on scores: (1) estimate person measures from the full set of items and then re-estimate them on the set of items with the misfitting items removed; (2) cross-plot the person measures from the two calibrations to determine their correlation; and, (3) determine if removing the misfitting items are consequential in terms of the decisions made (for example, the effect on classifying schools into three “performance” levels). If the cross-plot of these measures do not highlight any noticeable changes, then the “misfitting” items can be retained. The analyses at the school-level focused on whether the misfitting items bias the scores provided to schools and if schools were misclassified as a result (see section 5.5.2 for explanation of classification process). DESE uses a practical difference of 3 or more points in index scores at the school-level as a meaningful difference, and this difference was used as the criterion for assessing bias. One set of analyses focused on the removal of the five safety items (PSF7, BUL2, BUL5, BUL10, and BUL11) whose outfit mean square errors were above 2.0 (leaving 71 items in the analyses or it71); the other set examined the impact of removing all 14 misfitting items (Appendix E1: >1.5) on school-level parameter estimates (leaving 62 items in the analyses or it62).

The first set of analyses compared parameter estimates based on all 76 items with estimates from the 71-item calibration (it71). The second set of analyses compared parameter estimates of all 76 items with estimates from the 62-item calibration (it62). Each set of analyses was broken out by grade (scores were reported by grade to schools and districts). Schools who met the minimum reporting requirements (N of 10 and school-level reliability of 0.7) when all items (it76) were calibrated were used in these analyses. At the student level, when comparing the 76-item calibration to the 71-item calibration, the correlation between the parameter estimates was 0.99 for each grade; when comparing the 76-item calibration to the 62-item calibration, the individual-level correlations across the three grades were all 0.98. Parameter estimates were aggregated to the school-level to determine the impact of the misfitting items at this level. Table 4 shows the school-level score correlations for the two sets of calibrations.

Table 4:

School-level parameter correlations1

| **Calibration** | **G5 (N = 731)** | **G8 (N = 441)** | **G10 (N = 292)** |
| --- | --- | --- | --- |
| 76-item versus  71-item | 0.99 | 1.00 | 1.00 |
| 76-item versus  62-item | 0.98 | 0.99 | 0.99 |

1Correlations shown are for schools who met DESE’s minimum reporting requirements

Removing the five most misfitting items did not distort school-level estimates; there was a near perfect correlation between the two calibrations and no further analyses were performed. When all 14 misfitting items were removed, the correlations between calibrations were all close to 1. Further analyses were performed on this data. Seventeen schools’ average overall school climate scores differed by 3 or more points in grade 5 (2.3%), no schools differed in grade 8, and 4 schools differed by 3 or more points in grade 10 (1%). Of the schools that differed by 3 or more points, only 4 schools in grade 5 (0.5%) and 1 school in grade 10 (0.3%) were misclassified; these schools’ means were all centered at the cut points in each grade. Overall, the results indicate that the Rasch model was robust to the presence of item misfit and these items did not impact the meaning of the overall scores provided to schools and districts.

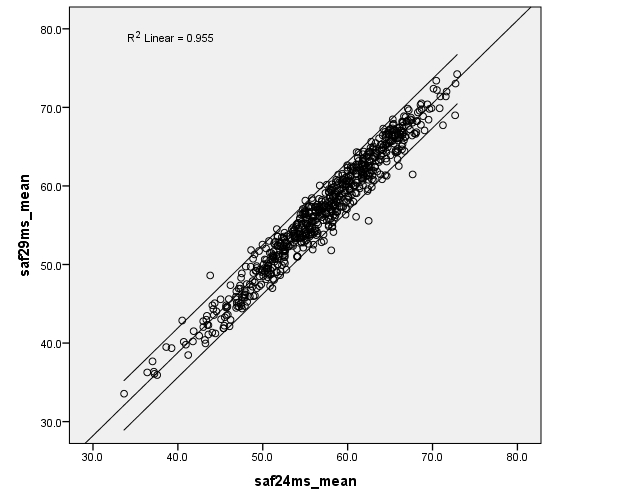
**Reliability.** When there are fewer items in the estimation process, the precision of the estimates can be affected, which in turn can reduce their reliability. A reduction in reliability could lead to fewer schools receiving school climate scores. When comparing the reliability of the scores provided to schools based on the 76-item calibration to the 62-item calibration, the same number of schools who served grade 5 and grade 10 received scores. One additional school serving grade 8 would have received a report when 62 items were used in the calibration. Inclusion of the misfitting items had a negligible impact on score reliabilities and the number of schools receiving a report.

**Subgroup scores.** The difference in mean student-level subgroup scores was assessed when scores were estimated using all 76 items and when using 62 items. There was no impact (means differed by less than 0.8 points) on subgroup scores broken out by gender, race/ethnicity, economically disadvantaged, and students with disabilities. English learner (EL) student scores in grade 5, grade 8, and grade 10 did differ by 1.6 points, 1.1 points, and 1.6 points, respectively. However, these EL differences are minimal and within the standard error of measurement. These data (bias, reliability, and subgroup analyses) suggest that there is no bias introduced when the misfitting items were retained in the calibration process.

* + 1. **Practical significance of misfitting items on safety scores.** Ten of the 14 misfitting items belonged to the safety items; the practical significance of these items on safety scores was assessed.When safety items were calibrated separately to estimate safety scores, only five of the 10 safety items still misfit the Rasch model (PSF7, BUL2, BUL5, BUL10, and BUL11). This calibration resulted in outfit mean squares for these items of between 1.5 and 2.0; this level of misfit should not degrade score measures (Linacre, 2019). In addition, their removal could potentially affect the number of schools receiving safety scores and threaten the content validity of the safety dimension. The practical significance of including the misfitting items in estimating individual, subgroup, and school-level safety measures, was evaluated. Two separate calibrations were run; one calibration included all 29 safety items (it29), with the other excluding the 5 misfitting items (it24). Analyses were again performed by grade and included schools who met the minimum reporting requirements (N of 10 and school-level reliability of 0.7) when all items (it29) were calibrated in the estimation process.

**School-level score bias.** At the student-level, the Pearson correlation between the two sets of measures was 0.98, 0.99, 0.99 for grade 5, grade 8, and grade 10 respectively. Upon aggregation, the correlation between it29 and it24 estimates was 0.98, 1.00, and 0.99 for grade 5, grade 8, and grade 10, respectfully. Figure 5 shows the correlation between school-level safety estimates for schools serving grade 5. In total, 15 schools (13, grade 5; 2, grade 10) had estimates that differed by 3 or more points when comparing the it29 calibration and the it24 calibration; however, only 2 of the schools serving grade 5 were misclassified. Retaining the misfitting safety items did not distort the safety measures at the school level and the safety measures are “theoretically and practically useful” (Crisan, Tendeiro, & Meijer, 2017, p. 453).

Figure 5. Correlation of G5 school safety estimates with (it29) and without (it24) misfitting items



**Reliability.** The removal of the misfitting items from the safety dimension estimation did have a negative impact on the number of schools who received a safety score; schools serving grade 5 were most affected. Seventy-two fewer schools (~12%) serving grade five would not receive a safety score if the misfitting items were removed. Nine fewer schools serving grade 10 (~3%) and 2 fewer (<1%) schools serving grade 8 would also not receive a safety score. The decrease in reliability due to the removal of misfitting items has a practical negative consequence on the reporting of safety dimension scores. A substantial proportion of schools and districts would not have access to safety scores that allow them to compare scores across dimensions, grades, schools, and time (2018 is the baseline year for trend data).

**Subgroup scores.** Average English learner scores differed marginally when safety scores from the 29-item calibration were compared to scores from the 24-item calibration. English learner scores differed by 1.3 and 1.2 points in grade 5 and grade 10, respectively (grade 8 EL scores did not differ). These data suggest that there is no bias introduced when all safety dimension items are used in the calibration process.

**5.1.4. Reverse-scored items and misfit.** When all 76 items were calibrated together, the primary source of misfit were reverse-scored items (11 of the 14 misfitting items were reverse-scored). Research has shown that reverse-scored items can be confusing to respondents and this is one reason why these types of items misfit the Rasch model (Conrad, Wright, McKnight, McFall, Fontana, and Rosenbeck, 2004). Additional analyses were performed to determine the suitability of including these items in the score estimation process. To determine whether students are confused by the 16 reverse-scored items (and by corollary their negative valence), all reverse-scored items were calibrated separately (Appendix E6). One item, SAFEMO11 (Because I worry about my grades, it is hard for me to enjoy school), had an outfit MNSQ error of 1.51; PTMs ranged from 0.53 to 0.70. The items explained 48% of the variance in students’ perceptions with no meaningful residual factor. The person separation reliability of the items ranges from 0.68 (real) to 0.73 (model). The item reliability was 1.00. These data suggest that students were not confused by the reverse-scored items and these items separated due to the “scoring method” (reversal). Evidence from the structural validity section (see section 5.2.1) indicates that these items did form a residual factor, but this component only explained 3.4 of the 121 observed variance units or 2.8% of the observed variance (well below Linacre’s criterion of 5% for multi-dimensionality).

Another test of whether students found the reverse-scored items confusing is to examine the category frequencies and observed step averages for each of the items; if item step averages do not increase monotonically, this suggests that students could have found the items confusing providing an explanation for these items misfitting the model. Appendix E7 shows the category frequencies and step averages for the reverse-scored items. All items except for PSF8 (Students are sexually harassed at my school (for example, bothered by unwanted touching and/or indecent name-calling)) have monotonically increasing observed step averages. This suggests that students are not confused by the content of these items and these results provide further support that the “scoring method” may lead to misfit and to the presence of a residual factor in the structural validity analyses.

* + 1. **Content validity conclusion.** The fit analyses support the use of the scores at the dimension level and for the bullying topic. Empirical analyses show that the inclusion of misfitting items does introduce some bias when the overall school climate scores are estimated; however, at the school level, the practical impact of this bias is minimal and nearly all schools are classified correctly. The misfit likely occurs due to the presence of the reverse-scored items forming a “scoring method” factor (Conrad, Wright, McKnight, McFall, Fontana, and Rosenbeck, 2004). Given the relatively low stakes attached to using the school climate scores (designed for use in continuous school improvement), the level of score bias and misclassification introduced was minimal and does not warrant the removal of these items from the survey.

The content of these items is particularly important to appropriately represent the safety dimension of the school climate construct. The bias created by including these misfitting items in estimating safety scores was again negligible. Removing the misfitting safety items did reduce the reliability of the safety measures at the school level; many schools, particularly those serving grade 5, would not have received a safety score were these items to be removed. Excluding these items would have had a real negative practical consequence with many schools not receiving a safety score. When combined with the evidence that the removal of the misfitting safety items did not significantly distort the measurement model, the misfitting safety items were retained. Item fit is one source of evidence to support the unidimensionality of the construct being measured by the Rasch model; another source is to access the dimensionality of the school climate construct using principal components analyses of the residuals.

* 1. **Structural validity**

Structural validity evaluates the alignment of the scoring structure to the hypothesized structure of the construct. The fundamental assumption of the Rasch model is that it measures only one latent construct (in this study, the school climate construct). If the data meet this assumption and other assumptions of the Rasch model, the measures are linear, invariant and additive; equal differences on the scale translate into equal differences in the probability of endorsing an item no matter where on the scale an item is located. In this validity study, the unidimensionality of the data was assessed by conducting (1) an assessment of the dimensionality data provided by the Rasch Winsteps software (Linacre, 2019), (2) an analysis of the standardized residuals, and (3) an examination of the correlational relationship between the freely calibrated dimension scores. These analyses were done for all 76 school climate items and separately for items belonging to each dimension (and bullying topic).

* + 1. **Overall dimensionality data (76 items).** Results from a principal component analyses of the residuals (Smith, 2002) using Linacre’s criteria (2019; Appendix D) for unidimensionality found that the variance explained by the 76-item measure was 37.1% (Table 5). The first contrast’s residual variance was less than 5% of the total observed variance. The variance explained by the items of the first dimension (school climate construct) is 6 times the variance explained by the first contrast (residual), meeting Linacre’s criterion of at least 4 times (Linacre, 2019).

Table 5

Residual analyses of 76-item VOCAL data (Grades 5, 8, and 10 combined)

| **Variance Component** | **Eigenvalue** | **Observed (%)** |
| --- | --- | --- |
| Raw variance explained by measures | 44.8 | 37.1 |
| Raw variance explained by persons | 25.9 | 21.5 |
| Raw Variance explained by items | 18.8 | 15.6 |
| Unexplained variance in 1st contrast | 3.4 | 2.8 |
| Item variance to 1st contrast multiple |  | 5.6x |

These data all support that the school climate construct is unidimensional. Although the residual variance was less than 5%, the eigenvalue was equal to 3.4, indicating the possibility of a second dimension. This was investigated.

* + 1. **Residuals Analyses of 76-item VOCAL data.** If the data fit the model and the variance in responses is explained by one latent trait (school climate construct), the unexplained or residual variance should be random (i.e., there is no relationship among the residuals). The Rasch dimensionality analyses first removes the common variance associated with all 76 items, and them examines the residuals. The residual analyses results are shown in Table 5 and Appendix F. The eigenvalue of 3.4 indicates that three to four items are forming an item cluster within the residuals. Linacre (Linacre, 2019, p. 544) reports that “in practice, we need at least 5 items to load heavily on a contrast, maybe more, before we consider those items as a separate instrument”. Five items have loadings above 0.4 on the 1st contrast; these items are related to bullying or aggressive student behaviors that can negatively impact school climate and all are reversed scored items (PSF 3, BUL5, BUL 13, BUL14, and BUL 15). The “scoring method” factor highlighted in the content validity section is leading to these items separating out in the dimensionality analyses. Except for BUL13 (a common item across all three grades), all first contrast items were on the grade 8 form of the VOCAL survey. Although these items are not “loading heavily” on the first contrast, the impact of these items on score estimation was assessed for each grade.

**First-contrast items.** When the five 1st contrast items were removed from the calibration, the remaining 71-item scale was of comparable reliability (real: 0.90; model: 0.92) to the 76-item scale (real: 0.91; model: 0.93). The variance explained increased slightly from 37.1% (it76) to 39.0% (it71); the first contrast explained only 2.5% of the observed variance and represented 2.9 of the 116 observed variance units. These data suggest that the 76-item scale was slightly more reliable than the 71-item scale. The minimal improvement of the 71-item scale in the unidimensionality data for the school climate construct did not warrant the removal of the five first contrast items. The impact of their removal on score estimation was also assessed.

The Pearson correlation of student-level measures between the 76-item scale and the 71-item scale was above 0.99 for students in each grade; of the schools with reportable data, the school-level correlation between mean scores was 1.0 for schools serving grades 5 and 10. The school-level correlation for schools serving grade 8 was 0.98; 1 of the 441 schools serving grade 8 would have been misclassified; this school’s overall school climate score was at the cut point between a “typical” school climate and a “relatively strong” school climate classification. Removal of the 1st contrast items made no impact on the score estimation process in grades 5 and 10, and negligible impact in grade 8.

* + 1. **Residual analyses of dimension/domain data.** When each dimension was calibrated and analyzed separately, the results supported the structural validity of each dimension; the residual analyses results are shown in Table 6. The variance explained in student perceptions were 41.5%, 38.7%, and 45.6% for the engagement, safety, and environment dimension, respectively, and 42.0% for the bullying domain. Of note, within the bullying climate domain, items that were designed to measure behaviors/practices that help protect students from bullying appeared to separate from items that measured actual bullying behaviors, and these items loaded on to the first contrast. However, the eigenvalue of the first contrast was only 2.0 (Table 6) indicating that items in the first contrast were not forming a second dimension.

Table 6

Residual analyses of dimension data (Grades 5, 8, and 10 combined)

| **Variance Component** | **Engagement**  **25 items** | **Safety**  **29 items1** | **Bullying**  **13 items1** | **Environment**  **22 items** |
| --- | --- | --- | --- | --- |
| Raw variance explained by measures | 41.5% | 38.7% | 42.0% | 45.6% |
| Raw variance explained by persons | 26.7% | 25.9% | 31.6% | 25.8% |
| Raw Variance explained by items | 15.3% | 12.8% | 10.4% | 19.8% |
| Unexplained variance in 1st contrast | 4.2% | 5.1% | 8.9% | 4.0% |
| Eigenvalue 1st contrast | 1.8 | 2.4 | 2.0 | 1.6 |
| Item variance to 1st contrast multiple | 3.6x | 2.5x | 1.2x | 4.9x |

1Bullying protective factor items (BUL1, BUL3, BUL4, and BUL9) separated from the bullying behavior items (BUL2, BUL5, BUL10 to BUL16)

The correlation between student estimates for the first residual cluster (bullying protective factors) and the other residual cluster (bullying behaviors) was 0.7; this indicates that the two clusters are related and measuring the bullying climate domain and no second dimension is distorting the measurement of this domain.

* + 1. **Sub-scale dimension/bullying correlations.** Student-level Pearson correlations were evaluated between sub-scale scores for the three separately calibrated dimensions of school climate (engagement, safety, and environment) and for bullying domain scores. The correlations should be positive and of sufficient magnitude (greater than 0.5 but less than 0.9) to indicate that the three sub-scales are measuring distinct but related dimensions of the school climate construct. The correlations were first estimated using all students in the analysis. Dimension subscale correlations ranged from 0.69 (safety and environment) to 0.80 (engagement and environment); the results are shown in Table 7 (below the diagonal).

The magnitude and pattern of correlations was also evident when examined for each grade separately (grade 5 data shown below the diagonal in Table 7). The lowest correlation (0.62) was between safety and environment scores in grade 10, the highest correlation (0.79) was between engagement and environment scores in grades 5 and 10. The overarching unifying construct of school climate explains the moderate-to-moderately strong relationship between the dimension scores.

When all students were included in the analyses, the correlations between bullying climate scores and the school climate dimension scores were 0.59, 0.92, and 0.54 for engagement, safety, and environment, respectively; the correlation between the bullying domain scores and overall school climate scores was 0.79. These moderate to strong correlations replicated across each grade (data not shown). These data support that the bullying domain items are theoretically related to each dimension and to the school climate construct overall.

Table 7

Pearson correlations between student dimension scores

| **Scale** | **Overall** | **Engagement** | **Safety** | **Environment** |
| --- | --- | --- | --- | --- |
| **Overall** | 1 | 0.90 | 0.88 | 0.89 |
| **Engagement** | 0.90 | 1 | 0.67 | 0.79 |
| **Safety** | 0.90 | 0.69 | 1 | 0.65 |
| **Environment** | 0.89 | 0.80 | 0.67 | 1 |

1Pearson correlations observed for all students are shown below the diagonal; grade 5 data above the diagonal

* + 1. **Structural validity conclusion.** The evidence from the dimensionality analyses, residual analyses, and the sub-scale correlational analyses supports the structural validity aspect of the school climate construct (76 items). The one dimension extracted by the Rasch model meets the unidimensionality assumption of the Rasch model, thereby supporting the use of scores for the intended purpose. The residual analyses highlighted that bullying behavior items separated from the other items; however, the signal-to-noise of this separation was not of sufficient magnitude to distort measures. The correlations of the sub-scale dimension scores (all students and by grade) support the theoretical premise that the school climate construct is composed of three related but distinct dimensions of school climate.
  1. **Substantive Validity**

Substantive validity assesses whether the responses to the items are consistent with the theoretical framework used to develop the items. Two sets of analyses are used to support the substantive validity aspect of construct validity: these are (1) an examination of the rating scale use by respondents, and (2) an assessment of whether the item difficulty hierarchy of the school climate survey conforms to best survey design principles (p. 4) and meets survey developers’ *a priori* expectations.

**Rating scale.** For each threshold of the rating scale, the mean square error fit statistics should be between 0.7 and 1.3. For surveys that use a four-point scale, the distance between the Andrich thresholds should be at least 0.8 logits (Appendix D; Wolfe & Smith, 2007b). In addition, the observed average for each response category should monotonically increase. The rating scale structure data and plot are shown in Figure 6.

Figure 6. Rating scale structure for VOCAL instrument

Figure 6 depicts the rating scale structure for the VOCAL instrument. Observed averages and Andrich thresholds increase monotonically. The plot shows that each response category (0,1,2,3) covers a distinct part of the construct continuum that ranges from -3.00 logits to +3.00 logits.

The rating scale for the 76 items of the VOCAL survey functioned well. Except for the little used score category of zero (never true), the category threshold fit statistics are excellent with MNSQ error near or equal to 1.00 (Figure 6). Adjacent Andrich category thresholds are greater than 0.8 logits apart and the observed average of each response category increases monotonically. Students are using the rating scale structure as intended. A qualitative assessment of how well the item difficulty hierarchy corresponds to the instrument developer’s *a priori* theoretical expectations also provides substantive validity evidence; this evidence is provided next for the overall school climate construct and for each dimension.

* + 1. **Overall VOCAL item hierarchy.** The overall item hierarchy across the school climate scale met DESE’s *a priori* expectations in terms of relative difficulty of individual items within and across dimensions. The ordered pattern of item difficulties also conforms to best test design principles (Wright and Stone, 1979). Figure 7 displays the item-variable map for the VOCAL survey with engagement items shown in yellow, safety items in pink, and environment items shown in green. Items for each dimension span the breadth of the school climate continuum with items from different dimensions overlapping as you move from low to high on the scale metric. Figure 8 shows the item threshold-variable map; calibrations cover approximately 98.4% of the student distribution. Some gaps in item calibrations are evident toward the top of the student distribution and at the bottom of the distribution. As a result, students at the tail ends of the distribution are measured with more error and are associated with larger standard errors. Appendix G provides the item difficulty hierarchy or measure order for all 76 items; item prompts are provided in Appendix H1 (engagement); H2 (safety), and H3 (environment) for reference.
    2. **Engagement dimension item hierarchy.** The ordered pattern of relative item difficulty within the relationship and cultural competence domains are consistent with expectations. For example, items related to student-on-student relationships (ENGREL1, ENGREL4, ENGCLC6, ENGCLC6) were, as expected, harder for students to affirm than items related to teacher-on-student relationships (ENGREL6, ENGREL6, ENGCLC1, ENGCLC2). Items that measure the degree to which classrooms are student-centered and integrate student ideas and interests into the structure of lessons (ENGPAR4, ENGPAR10) were the hardest for students to affirm. Similarly, items related to providing students with a degree of choice in their learning (ENGPAR5, ENGPAR11) were easier for students to endorse when compared to the student-centered items but were still relatively hard for students to affirm. Items related to participation in school life (PAR1, PAR2, PAR3) were relatively easy for students to endorse. These findings were as expected and are consistent with past research (Thomas, 2004; Peoples, O’Dwyer, Wang, Brown, & Rosca, 2014).

**Safety dimension item hierarchy.** The relative ordering of items within the safety dimension met prior expectations. Items related to students’ physical safety (e.g., SAFPSF2, SAFPSF4, and SAFPSF8) were relatively easy to disaffirm (a positive outcome) compared to items related to students’ emotional safety (e.g., SAFEMO1, SAFEMO8, SAFEMO9) or to items related to bullying protective behaviors (e.g., SAFBUL1, SAFBUL3, and SAFBUL4). Physical safety is a harbinger of a positive school environment and it was expected that physical safety items would be among the easiest to disaffirm. In contrast, emotional safety items were, as predicted, among the most difficult items on each grade’s survey for students to affirm. For students to reach out for emotional support from their teachers (SAFEMO 4, SAFEMO8) or from their peers (SAFEMO10) requires a complex interplay of students’ and teachers’ social and emotional competence; healthy teacher-student relationships are fundamental to positive school and classroom climates and are a cornerstone of effective classroom management (Jennings and Greenberg, 2009). As expected, these types of items are among the most difficult for students to affirm within the safety dimension.

Figure 7. Item-variable map for VOCAL survey items (Engagement items are in yellow; safety items are in pink, and environment items are in green.)

Figure 7 depicts the item-variable map for the VOCAL survey items. Engagement items are in yellow; safety items are in pink, and environment items are in green. Items and students (hatchets) are placed on the same scale metric that ranges from -5.00 logits to +6 logits.

Figure 8. Item-threshold map for VOCAL

Figure 8 shows the item-threshold map for the VOCAL item step thresholds. Item threshold parameters range from approximately -3.00 logits to +4 logits.

The item hierarchy within safety topics also met DESE’s expectations. For example, the three bullying items that asked about students’ perceptions of how well adults intervene to prevent bullying were easier for students to affirm than the one item that asked students if students intervened to stop bullying. In more than 80% of bullying situations, students take on bystander roles (assist or reinforce the aggressor, ignore the situation, or try to prevent); even when peers are present, approximately 1 in 3 students express being bullied in the previous two months (Polanin, Espelage and Pigott, 2010). To create the conditions in schools where student prosocial bystander behavior is encouraged and expected requires bullying prevention programs targeting these behaviors and strong supportive school climate (Polanin, Espelage and Pigott, 2010; Johnson, Waasdorp, Debnam and Bradshaw, 2013).

* + 1. **Environment dimension item hierarchy.** Item hierarchies in both discipline and mental health environments met *a priori* expectations. For example, items (DIS1, DIS7) that provide students with a voice in deciding school rules or consequences for poor behavior are harder to endorse than those that ask students about the fairness or consistency of enforcing school rules (DIS2, DIS4). Similarly, mental health items that rely on students’ awareness or management of their emotions (MEN9, MEN7, MEN4) were easier to affirm than items that related to whether the schools have “systems” developed to support students (MEN3, MEN6).

Table 8 provides a specific example of the item hierarchy from the instructional environment domain. Foundational to a positive instructional environment is the perception that teachers support and believe all students can succeed academically and set high expectations for student learning (TNTP, 2018). Items, such as, “My teachers believe that all student can do well in their learning (ENVINS8) and “My teachers set high expectations for my work” (ENVINSS5) were, as expected, relatively easy for students to endorse. In contrast, instructional environments that are collaborative, challenging, and relevant are much harder to engender (Peoples, Abbott, and Flanagan, 2015a, 2015b); these items were among the most difficult for students to affirm. This ordered pattern of item difficulties confirms developers’ *a priori* expectations.

Table 8.

Item hierarchy of instructional environment items

| **Item code** | **Grade** | **Item Prompt** | **Item Difficulty (logits)** |
| --- | --- | --- | --- |
| ENVINS14 | 5 | When I am home, I like to learn more about the things we are learning in school. | 1.64 |
| ENVINS12 | 10 | The things I am learning in school are relevant (important) to me. | 0.79 |
| ENVINS13 | 10 | Teachers ask students for feedback on their classroom instruction. | 0.76 |
| ENVINS 9 | 5,8 | My school work is challenging (hard) but not too difficult. | 0.44 |
| ENVINS1 | 5,8,10 | Students help each other learn without having to be asked by the teacher. | 0.22 |
| ENVINS11 | 5,8,10 | My teachers support me even when my work is not my best. | -0.13 |
| ENVINS15 | 10 | My teachers inspire confidence in my ability to be ready for college or career. | -0.25 |
| ENVINS5 | 8,10 | My teachers set high expectations for my work. | -0.64 |
| ENVINS2 | 5,8 | My teachers are proud of me when I work hard in school. | -1.03 |
| ENVINS3 | 5 | My teachers help me succeed with my school work when I need help. | -1.15 |
| ENVINS8 | 8 | My teachers believe that **all** students can do well in their learning. | -1.36 |

* + 1. **Substantive validity conclusion.** The well-functioning rating scale combined with the theoretically grounded 76-item item hierarchy provides the evidence needed to support the substantive validity aspect of the school climate construct. Items for each dimension are sufficiently dispersed along the school climate continuum and cover the target distribution well. Because of this coverage, most students are measured with minimum error for each of the three dimensions and for the school climate construct overall.
  1. **Generalizability**

A measure is considered generalizable when the score meaning and properties function similarly across multiple contexts (e.g., grades, subgroups, forms) or time points. Reliability analyses and differential item functioning (DIF) analyses are used to assess the generalizability of the measures. Similar to Cronbach’s alpha, person separation reliability (PSR) looks at the stability (internal consistency) of the measures across each of the forms and scoring structures. The reliability indices depict the ratio of true variance to observed variance; in the Rasch model, the person separation reliability index measures the ratio of the variance in latent person measures to the estimated person measures (Schumacker and Smith, 2007). Unlike classically derived measures, reliability estimates are available for items as well as for persons using Rasch methodology. Standard errors are estimated for each person and each item and are used to provide an estimate of error variance (Schumacker and Smith, 2007). DESE used DIF analyses to empirically test for item invariance across several subgroups; item invariance ensures comparability of score interpretation.

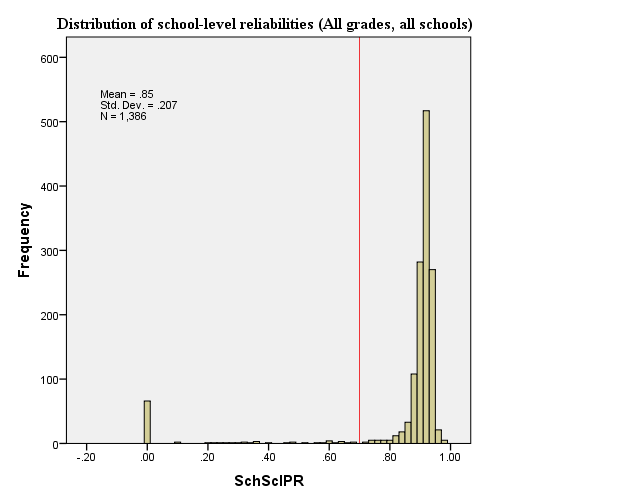
* + 1. **Reliability evidence:** Best test design principles (Wright, 1979) necessitate the alignment of the mean of the item distribution to the mean of the person distribution. The mean person difficulty of the 76-item scale was +1.06 logits with a standard deviation of 0.99 logits (Appendix I). The items are reasonably well targeted for the student distribution (Figure 7; Appendix I) resulting in a real person separation reliability (PSR) of 0.91, and a real person separation index of 3.11. Notable in Figure 7 is the relative rarity of some bullying and physically aggressive behaviors when compared to other indicators assessed; these off-target items likely contribute to the misalignment of the person and item distributions. The real person separation reliabilities were: 0.90 for the 36-item grade 5 form; 0.90 for the 38-item grade 8 form, and 0.89 for the 38-item grade 10 form (Appendix I). The replication of reliabilities across forms provides evidence for the reproducibility and stability of the school climate construct. Reliabilities above 0.8 are considered acceptable for the current use of the surveys (Appendix D), namely to provide schools and districts with formative data to use for continuous improvement. New items will be tried out in the 2019 VOCAL administration with the goal of improving the reliability of each grade-level survey.

Appendix I shows the reliability of each dimension when the three grades are calibrated together. The real person separation reliability of the engagement, safety, and environment scores was 0.77, 0.81, and 0.76, respectively. These reliabilities are likely attenuated due to the design of the test forms (Schwartz, Ayers, and Wilson, 2017). Students across the three grades only responded to a small sub-set of common items for each dimension; students largely responded to sets of unique items. As a result, a large amount of “missing data” is realized when the three grades’ data were combined to assess the reliability of each dimension. The true reliabilities of the dimension scores are underestimated (Schwartz, Ayers, and Wilson, 2017).

**School-level reliability.** In reporting out climate scores to schools, it is important to ensure that schools receive reliable data. On average, of the districts and schools who participated, ninety-one percent of districts and eighty-one percent of schools received VOCAL scores, respectively. For a school to receive an index score, 10 or more students had to contribute to the score and the school-level person separation reliability of each index score had to be 0.7 or more. Most schools who did not receive an overall or dimension score did not have enough students to receive a report.

Figure 9 shows the distribution of the overall school climate index reliabilities within the sample. The average reliability across the 1,386 schools in the sample was 0.85 and reliabilities ranged from 0 to 0.97. By grade, 91%, 84%, and 72% of schools serving grade 5, grade 8, and grade 10, respectively, met the minimum reliability requirement. By dimension, 77% of schools met the minimum reliability requirement for an engagement score; 85% for a safety score; and 78% for an environment score.

Figure 9. Distribution of school-level school climate score reliabilities



* + 1. **Differential item functioning (DIF) analyses**

To support the claim that the school climate instrument is generalizable, the items should have the same meaning for different subgroups of respondents (e.g., gender, race/ethnicity). Respondents of the same ability (endorsement level), should have the same probability of affirming an item irrespective of the subgroup they belong to. In this study, items were flagged if they average difficulties differed by 0.5 logits or more (Appendix D). The analyses indicated that item deltas did not differ significantly across the following subgroups: gender, race, students with disabilities, and economically disadvantaged; over 90% of items differed by less than 0.5 logits. One engagement item (CLC4, administered in G10) exhibited severe DIF (>1.0 logit) when comparing students with disabilities to students without disabilities. Similarly, one safety item (BUL11, administered in G10) exhibited severe DIF (>1.0 logit) when comparing white students to all other racial/ethnic subgroups.

DIF was present for English learners with twelve items having DIF of greater than 0.7 logits. Seven of the twelve items (PAR3, PAR12, EMO6, BUL10, BUL11, BUL16, PSF5, INS12) were on the grade 10 form. Four of the remaining five items (PAR1, PAR7, PSF7, BUL12) were administered on the grade 5 form; one item, PSF4, exhibited DIF on the grade 8 form. Seven of the twelve DIF items were structured as negative valence items and required reverse scoring; eight of the twelve items were from the safety dimension and four from the participation topic within the engagement dimension. Two of the twelve items that displayed severe DIF across EL groups also exhibited DIF across some, but not all, race groups (BUL11, PSF7). White students found these two items easier to disaffirm than non-white students. DESE’s surveys were not translated for English learners. Language barriers likely explain the DIF present across certain race/ethnicity and EL subgroups with students unable to properly access the survey content. Figure 10 and Figure 11 show DIF plots for gender and race status, respectively. DIF plots for the remaining subgroup comparisons are found in Appendix J1 (economically disadvantaged), J2 (students with disabilities), and J3 (English learner), respectively.

**EL students and DIF.** Analyses were performed to determine the impact of including the twelve DIF items in the EL students’ overall score estimations. Subgroup data are not reported out at the dimension level. EL students’ overall school climate scores were estimated with and without these DIF items included. Grade-level analyses focused on how many schools would be mis-classified if these DIF items were included in the EL subgroup score estimation. Based on schools who met the minimum reporting criteria, 5.2% (38 schools) serving grade 5 would be mis-classified; correspondingly, 1.6% (7 schools) and 3.8% (11 schools) of schools serving grade 8 and grade 10 would be mis-classified, respectively.

If scores are sufficiently reliable, EL dimension scores may be reported out to schools in the future. The majority of the DIF items were from the safety dimension; the analyses above were performed using only safety items. 9.8% (79 schools) of schools serving grade 5 would be mis-classified for EL subgroup scores; correspondingly, 2.5% (11 schools) and 8.2% (24 schools) of schools serving grade 8 and grade 10 would be mis-classified, respectively.

The decision was made to remove the 12 items from the overall school climate calibration process when estimating EL students’ subgroup scores; in total, 64 items of the76 items were used to estimate EL students’ scores. Because none or only 1 – 2 items exhibited DIF in other subgroup comparisons (gender, economically disadvantaged, race, and student with disabilities), these items were retained when reporting out these sub-group scores.

Figure 10. Differential item function plot by gender

Figure 11. Differential item function plot by race/ethnicity

* + 1. **Generalizability conclusion**

The reliability data for the overall school climate scale and the reliability data for the three dimensions support the generalizability of the construct and associated dimensions. The

majority of items exhibited no DIF across five different subgroup comparisons. *Scores for EL students should be viewed with caution due to the decreased number of items used to estimate EL subgroup scores.*

* 1. **External validity**

This aspect of construct validity relates to the responsiveness of an instrument and the relationship of its scores to the scores of external measures (criterion validity). The responsiveness of an instrument refers to “the degree to which an instrument is capable of detecting changes in person measures following an intervention that is assumed to impact the target construct” (Wolfe & Smith, 2007b, p. 222). If an instrument is responsive, it can be applied appropriately to measure expected group differences or individual/group change. The first section (5.5.1) examines the instrument’s responsiveness at the student-level; the second section (5.5.2) assesses responsiveness at the school-level and its impact on reportable scores.

Criterion validity is the strongest form of external validity; it determines how well scores from an instrument predict scores on a criterion measure (e.g., how well do school climate scores predict achievement). There are two forms of criterion validity: concurrent and predictive. This section reports data to support the concurrent validity of the VOCAL survey scores. Because the unit of interest is the school, the external validity analyses focus on examining the relationship between both school-level aggregate VOCAL scores and school-level aggregate scores of the following criterions: student achievement, attendance, chronic absence, discipline rates, suspension rates, and retention rates. Concurrent validity is discussed in section 5.5.3.

* + 1. **Student-level responsiveness**. The responsiveness of an instrument is measured by the person strata index, H, which provides the number of statistically distinct endorsement groups whose centers of score distributions are separated by at least three standard errors of measurement within the sample. According to the formula, H = (4G +1)/3 (Wright and Masters, 2002, p. 888) and a real person separation index, (PSI; G) of 3.1, the number of person strata for the 76-item VOCAL instrument is equivalent to almost 4.5 distinct person strata (Appendix I). The number of person strata ranged from 4.2 in grades 5 and 10 to 4.4 in grade 8. The VOCAL instrument produces reliable, reproducible measures which are responsive (i.e., the instrument can divide the sample into four to five statistically distinct score groups).
    2. **School-level responsiveness and score reporting.** The greater the number of person strata at the individual-level, the more likely the instrument will be able to meaningfully differentiate schools. At the school-level, the average scaled score was 1.18 logits with a standard deviation of 0.43 logits (Table 9). After removing schools whose data did not meet DESE’s minimum reporting requirements (N of 10 and school-level person separation reliability of at least 0.7), reportable school measures ranged from ‑0.10 logits to 2.60 logits indicating variability in school-level scores. Table 9 shows the highest and lowest school scores broken out by school type (elementary, middle, and high school); the scores shown are for schools with response rates above 85%. The relatively high degree of responsiveness of the instrument at the student-level appears to pick up the variation within and between schools.

Table 9

Variability of reportable school-level VOCAL scores

|  | **Response rate** | **Person Separation Reliability (PSR)1** | **Mean ±SD2**  **Logits** | **Mean Transformed** |
| --- | --- | --- | --- | --- |
| Weaker Elementary | 96% | 0.94 | 0.28 ± 1.04 | 35 |
| Weaker Middle | 91% | 0.93 | 0.22 ± 1.08 | 33 |
| Weaker High | 88% | 0.92 | 0.28 ± 0.86 | 34 |
| **Average school** | **-----** | **-----** | **1.18 ± 0.43** | **52** |
| Stronger Elementary | 100% | 0.85 | 2.60 ± 1.19 | 78 |
| Stronger Middle | 90% | 0.90 | 1.70 ± 0.98 | 62 |
| Stronger High | 94% | 0.91 | 1.50 ± 1.02 | 58 |

1A PSR of 0.7, and a N of 10 or more students was set as the minimum reporting requirements 2SD: Standard Deviation

**Score reporting and profiles.** DESE linearly transformed logit measures to make them more interpretable. The student-level logit measures were standardized and the zscores transformed to have a mean of 50 and a standard deviation of 20 (see Appendix K for details). The individual scores were truncated and placed on a scale of 1 – 99 (± 2.5 standard deviations) and then aggregated up to the school level. School-level scores had a mean of 50.05 and a standard deviation of 12.83; schools with reportable data had an average score of 52.4 and a standard deviation of 8.5. To help schools interpret their data in each grade, schools were separated into three “performance” levels using the mean and standard deviation of the school-level scores. Based on the median student within these three “performance” groups and all available items, a profile or picture of the school climate in each group was constructed using the probabilities of the median student responding in each of the four response categories. These profiles are provided in Table 10a (grade 5), 10b (grade 8), and 10c (grade 10).

|  |  |  |
| --- | --- | --- |
| **Table 10a: -\_Grade 5 VOCAL profile: Statewide, scores range from 33 to 78** | | |
| **Schools with relatively weak school climates (bottom 15% of schools)** | **Schools with typical school climates**  **(middle 70% of schools)** | **Schools with relatively strong school climates (top 15% of schools)** |
| **33 to 51** | **52 to 65** | **66 to 78** |
| 1. Student-on-student relationships are largely respectful and caring but students are less open to having inclusive relationships with a diverse range of students. Adults model caring and respectful interactions. 2. Teacher expectations for student effort and perseverance are less demanding. Teachers’ help students succeed academically. Some students may benefit from more adaptive explanations to understand and access content. Teachers use student ideas, interests, and sharing to help students learn. The classroom environment is collaborative and supportive among students, and between students and teachers. Most students view their school work as appropriately challenging but tend not to want to or are unable to learn more when home. 3. Students feel fairly safe in school. Bullying behaviors are present. Teachers/adults try to counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. Not all students feel school rules are fair for all students. Students have limited say in deciding these rules. 4. Students, overall, feel happy in school but have a more moderate sense of belonging to their school. Schools teach students how to develop caring relationships and how to manage their emotions when angry or upset. Teachers reach out to help distressed students with most students feeling comfortable seeking help. Students can also largely rely on their peers for emotional support. | 1. Student-on-student relationships are largely respectful and caring with students open to having inclusive relationships with a diverse range of students. Adults actively model caring and respectful interactions. 2. Teachers have high expectations for student effort and perseverance. Teachers help students succeed academically by using different strategies to explain and make content accessible. Teachers use student ideas, interests, and sharing to help students learn. The classroom environment is collaborative and supportive among students, and between students and teachers. Most students view their school work as appropriately challenging and enjoy learning more when home. 3. Students feel safe in school. Bullying behaviors are present but teachers/adults actively counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. Not all students feel school rules are fair for all students. Students have limited say in deciding these rules. 4. Students, overall, feel happy in school but have a more moderate sense of belonging to their school. Schools teach students how to develop caring relationships and how to manage their emotions when angry or upset. Teachers reach out to help distressed students with most students feeling comfortable seeking help. Students can also largely rely on their peers for emotional support. | 1. Student-on-student relationships are largely respectful and caring with students open to having inclusive relationships with a diverse range of students. Adults actively model caring and respectful interactions. 2. Teachers have high expectations for student effort and perseverance. Teachers help students succeed academically by using different strategies to explain and make content accessible. Teachers use student ideas, interests, and sharing to help students learn. The classroom environment is very collaborative and supportive among students, and between students and teachers. Most students view their school work as appropriately challenging and enjoy learning more when home. 3. Students feel very safe in school. Some bullying behaviors are present but teachers/adults actively counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. A large majority of students feel school rules are fair for all students. Students have a say in deciding these rules. 4. Students, overall, feel very happy in school and have a strong sense of belonging to their school. Schools actively teach students how to develop caring relationships and how to manage their emotions when angry or upset. Similarly, teachers actively reach out to help distressed students. As a result, students feel comfortable seeking help. Students can also largely rely on their peers for emotional support. |
| The average student within these schools responds, “mostly true” to a large majority of items, and “always true” and “mostly untrue” to three and two items. | The average student within these schools responds “mostly true” to most items, “always true” to all but one of the remaining items, and “mostly untrue” to one item. | The average student within these schools responds “always true” to a majority of items, and “mostly true” to all remaining items. |

|  |  |  |
| --- | --- | --- |
| **Table 10b: - Grade 8 VOCAL profile: Statewide, scores range from 32 to 70** | | |
| **Schools with relatively weak school climates (bottom 15% of schools)** | **Schools with typical school climates**  **(middle 70% of schools)** | **Schools with relatively strong school climates (top 15% of schools)** |
| **32 to 41** | **42 to 50** | **51 to 70** |
| 1. Student-on-student relationships lack respect with students less open to having inclusive relationships with a diverse range of students. Adults generally promote and model respectful interactions among and between students, and with students’ families. 2. Teachers set moderately high expectations and are available when students need help. Teachers encourage students to work hard and try to instill a belief that all students can do well. Teachers tend not to use student ideas, cultural backgrounds, and interests to plan and guide their instruction, or to provide students with a choice in how to show their learning. Most students view their school work as appropriately challenging. The classroom environment is predominantly collaborative and supportive among students and between students and teachers.      1. Students feel fairly safe in school. Bullying behaviors are more prevalent. Teachers/adults try to counteract these behaviors. Students will largely not intervene to prevent bullying. When students are in trouble, students generally are not provided with a chance to explain their behavior. To reduce behavioral problems, students are taught how to settle conflicts by themselves. Staff are generally consistent when enforcing rules, but students express having no say in deciding these rules. 2. Students feel stressed about their grades. Most students have access to relatively effective social and emotional support systems. Despite believing their teachers are interested in their emotional well-being and teachers are trying to reach out to help distressed students, students feel relatively uncomfortable approaching teachers and counselors for help. Students are less able to rely on their peers for support when they are upset. | 1. Student-on-student relationships are largely respectful with students open to having inclusive relationships with a diverse range of students. Adults generally promote and model respectful interactions among and between students, and with students’ families. 2. Teachers set moderately high expectations and are available when students need help. Teachers actively encourage students to work hard and instill a belief that all students can do well. Teachers use student ideas, cultural backgrounds, and interests to plan and guide their instruction and do allow students to choose how they want to show their learning. Most students view their school work as appropriately challenging. The classroom environment is predominantly collaborative and supportive among students and between students and teachers. 3. Students feel safe in school. Bullying behaviors are present but teachers/adults try to counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. To reduce behavioral problems, students are taught how to settle conflicts by themselves. Staff are generally consistent when enforcing rules, but students express having limited say in deciding these rules. 4. Students feel stressed about their grades. Most students have access to relatively effective social and emotional support systems. Because most students believe their teachers are interested in their emotional well-being and teachers try to reach out to help distressed students, students feel relatively comfortable approaching teachers and counselors for help. Most students can also rely on their peers for emotional support when they are upset. | 1. Student-on-student relationships are largely respectful with students open to having inclusive relationships with a diverse range of students. Adults actively promote and model respectful interactions among and between students, and with students’ families. 2. Teachers set high expectations and are readily available when students need help. Teachers actively encourage students to work hard and instill a belief that all students can do well. Teachers use student ideas, cultural backgrounds, and interests to plan and guide their instruction and do allow students to choose how they want to show their learning. Most students view their school work as appropriately challenging. The classroom environment is predominantly collaborative and supportive among students and between students and teachers. 3. Students feel safe in school. Bullying behaviors are present but teachers/adults actively counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. To reduce behavioral problems, students are taught how to settle conflicts by themselves. Staff are generally consistent when enforcing rules, but students express having limited say in deciding these rules. 4. Students feel some stress about their grades. Most students have access to relatively effective social and emotional support systems. Because most students believe their teachers are interested in their emotional well-being and teachers try to reach out to help distressed students, students feel relatively comfortable approaching teachers and counselors for help. Most students can also rely on their peers for emotional support when they are upset. |
| The average student within these schools responds “mostly true” to most items, “mostly untrue” to all but one of the remaining items, and “never true” to one item, respectively. | The average student within these schools responds, “mostly true” to a large majority of items and “always true” and “mostly untrue” to four and two items, respectively | The average student within these schools responds “mostly true” to most items, “always true” to all but one of the remaining items, and “mostly untrue” to one item. |

|  |  |  |
| --- | --- | --- |
| **Table 10c: - Grade 10 VOCAL profile: Statewide, scores range from 27 to 67** | | |
| **Schools with relatively weak school climates (bottom 15% of schools)** | **Schools with typical school climates**  **(middle 70% of schools)** | **Schools with relatively strong school climates (top 15% of schools)** |
| **27 to 41** | **42 to 50** | **51 to 67** |
| 1. Student-on-student relationships lack respect with students less open to having inclusive relationships with a diverse range of students. Adults generally promote and model respectful interactions among and between students. 2. Teachers set moderately high expectations and are available when students need help. Teachers tend not to use student feedback, ideas, or interests to guide their instruction. Students view their learning as relatively irrelevant. Encouragement and opportunities for students to challenge themselves to learn are largely limited. The classroom environment is predominantly collaborative and supportive among students and between students and teachers. Teachers generally inspire confidence in students’ ability to succeed after high school.      1. Students feel fairly safe in school. Bullying behaviors are more prevalent. Teachers/adults try to counteract these behaviors. Students will largely not intervene to prevent bullying. When students are in trouble, most students are not provided with a chance to explain their behavior. Any disciplinary consequences are generally consistent across all students. Students have no say in deciding school rules. 2. Students feel stressed about their grades and most consider the level of academic pressure somewhat unhealthy. Students are less able to rely on their friends to help them cope with any emotional problems, or supportive friendships are missing. Most students report having access to relatively effective social and emotional support systems. Teachers, for the most part, reach out to help students emotionally. Students have a more moderate sense of belonging to their school. | 1. Student-on-student relationships are largely respectful with students open to having inclusive relationships with a diverse range of students. Adults generally promote and model respectful interactions among and between students. 2. Teachers set moderately high expectations and are available when students need help. Teachers use student feedback, ideas, and interests to guide their instruction. Students view their learning as mostly relevant. Encouragement and opportunities for students to challenge themselves to learn are largely available. The classroom environment is predominantly collaborative and supportive among students and between students and teachers. Teachers generally inspire confidence in students’ ability to succeed after high school. 3. Students feel safe in school. Some bullying behaviors do occur, but teachers/adults try to counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. Any disciplinary consequences are generally consistent across all students. Students have limited say in deciding school rules. 4. Students feel stressed about their grades, but most do not consider the level of academic pressure unhealthy. They rely heavily on their friends to help them cope with any emotional problems. Most students report having access to relatively effective social and emotional support systems. Teachers, for the most part, reach out to help students emotionally. Students have a strong sense of belonging to their school. | 1. Student-on-student relationships are largely respectful with students open to having inclusive relationships with a diverse range of students. Adults actively promote and model respectful interactions among and between students. 2. Teachers set high expectations and are readily available when students need help. Teachers use student feedback, ideas, and interests to guide their instruction. Students view their learning as mostly relevant. Encouragement and opportunities for students to challenge themselves to learn are largely available. The classroom environment is predominantly collaborative and supportive among students and between students and teachers. Teachers generally inspire confidence in students’ ability to succeed after high school. 3. Students feel very safe in school. Some bullying behaviors do occur, but teachers/adults actively counteract these behaviors. Students also try to prevent bullying. When students are in trouble, most students are given a chance to explain their behavior. Any disciplinary consequences are generally consistent across all students. Students have limited say in deciding school rules. 4. Students feel some stress about their grades, but most do not consider the level of academic pressure unhealthy. They rely heavily on their friends to help them cope with any emotional problems. Students have ready access to effective social and emotional support systems. Teachers, for the most part, reach out to help students emotionally. Students have a strong sense of belonging to their school. |
| The average student within these schools responds “mostly true” to most items, “mostly untrue” to all but one of the remaining items, and “never true” to one item. | The average student within these schools responds, “mostly true” to a large majority of items and “always true” and “mostly untrue” to four and two items, respectively | The average student within these schools responds “mostly true” to most items, “always true” to all but one of the remaining items, and “mostly untrue” to one item. |

Relatively weaker schools had scores that fell 1 or more standard deviations below the grade-level mean; relatively stronger schools had scores that fell 1 or more standard deviations above the grade-level mean. Approximately fifteen percent of the schools with reportable data were assigned to the top (stronger climate) or bottom (weaker climate) “performance” level; approximately seventy percent of schools were characterized as “typical”.

The VOCAL survey meaningfully differentiated schools both quantitatively and qualitatively. The profiles were designed to help schools assess and improve their climates. For schools that fall within the “weak” category, the profile provides them with a starting point to begin their analyses of student perceptions. For example, students in schools with relatively weak school climates report that students are not respectful or caring; in contrast students within schools with relatively strong school climates report that student-on-student relationships are respectful, caring and collaborative. These profiles offer a broad, relatively coarse guide to improvement; individual schools can use DESE’s [analytical and planning tool](http://www.doe.mass.edu/research/vocal/2018/interpretive-guide.docx) to get a more in-depth understanding of students’ perceptions within their schools.

* + 1. **Concurrent Validity.** Preliminary evidence of concurrent validity at the school level indicates a correlational relationship between students’ overall school climate scaled scores and several school-level criteria. When all schools are examined together, there is a small to moderate statistically significant relationship between VOCAL scaled-scores and attendance rates (0.32), chronic absence rates (-0.34), disciplinary rates (‑0.51), in-school suspension rates, and (-0.34), out-of-school suspension rates (‑0.34). These patterns of associations were reproduced across the three grades in 2018 and also replicated across years ([DESE, 2018](http://www.doe.mass.edu/research/vocal/2017/validity-study.docx), p.34). In addition, within high schools, graduation rates (0.12) and dropout rates (-0.08) were related to school climate scores; although statistically significant, the magnitude of these correlations was small. All these associations were in the expected direction; these data are summarized in Table 11.

Table 11

School-level correlations of 2018 criterion indicators and overall VOCAL scores1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade** | **All Schools**  **(N = 1,227) 9** | **Grade 5**  **(N = 731) 9** | **Grade 8**  **(N = 441) 9** | **Grade 10**  **(N = 292) 9** |
| Attendance rate2 | 0.32\*\*\* | 0.20\*\*\* | 0.17\*\*\* | 0.14\* |
| Chronically absent3  (10% or more) | -0.34\*\*\* | -0.25\*\*\* | -0.20\*\*\* | -0.15\* |
| Discipline rate4 | -0.51\*\*\* | -0.37\*\*\* | -0.34\*\*\* | -0.47\*\*\* |
| In-school suspension (ISS) 5 | -0.34\*\*\* | -0.19\*\* | -0.21\*\*\* | -0.34\*\*\* |
| Out-school suspension (OSS) 6 | -0.34\*\*\* | -0.35\*\*\* | -0.28\*\*\* | -0.34\*\*\* |
| Graduation rate7 | NA | NA | NA | 0.12\*\*\* |
| Drop-out rate8 | NA | NA | NA | -0.08\*\*\* |

1Data based on schools with greater than or equal to 10 students contributing to the aggregate VOCAL score, and a minimum school-level reliability of 0.7; 2Attendance rate: Indicates the average percentage of days in attendance for students enrolled in grades PK–12; 3Chronically absent (10% or more): The percentage of students who were absent 10% or more of their total number of student days of membership in a school. 4Discipline rate: the number of disciplinary incidents divided by school enrollment; 5In-School Suspension Rate: The percentage of enrolled students in grades 1–SP who received one or more in-school suspensions. 6Out-of-School Suspension Rate: The percentage of enrolled students in grades 1–SP who received one or more out-of-school suspensions; 7Graduation rate: The percentage of students who enroll in high school and graduate within 4 years, N = 268; 8Drop-out rate: The percentage of students in grades 9-12 who dropped out of school between July 1 and June 30 prior to the listed year and who did not return to school by the following October 1, N = 268. ; 9\*\*\**p*<0.001; \*\**p*<0.01; \* *p*<0.05

A positive statistically significant relationship between students’ VOCAL scaled scores and achievement at the school level was found. The Pearson correlations between the Massachusetts Comprehensive Assessment (MCAS) English Language Arts and Literacy (ELA) scores, mathematics scores, and Science and Technology/Engineering (STE) scores were 0.46, 0.45, and 0.20, respectively (Table 12). School-level VOCAL scores were also positively related to students’ aggregate growth percentile scores in ELA (0.26) and mathematics (0.25). These significant associations between school climate and achievement scores replicated across grade 5, grade 8, and grade 10; however, the magnitude of the relationships declined within high schools.

Table 12

School-level correlations of 2018 achievement scores and overall VOCAL scoresl1,2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade** | **All Schools**  **(N = 1,227)3** | **Grade 5**  **(N = 731)3** | **Grade 8**  **(N = 441)3** | **Grade 10**  **(N = 292)3** |
| English Language Arts and Literacy scaled score | 0.46\*\*\* | 0.32\*\*\* | 0.27\*\* | 0.12\* |
| English Language Arts and Literacy student growth percentile | 0.26\*\*\* | 0.32\*\*\* | 0.16\*\* | 0.10 |
| Mathematics scaled score | 0.45\*\*\* | 0.36\*\*\* | 0.27\*\* | 0.16\* |
| Mathematics student growth percentile | 0.25\*\*\* | 0.30\*\*\* | 0.16\*\* | 0.19 |
| Science and Technology/Engineering scaled score4 | 0.20\*\*\* | 0.31\*\*\* | 0.26\*\* | 0.15\* |

1Data based on schools with greater than 10 students contributing to the aggregate VOCAL score, and had a school-level VOCAL reliability of 0.7 or more; 2Grade 5 and grade 8 MCAS tests reflect DESE’s new generation assessments; the grade 10 test is based on the old legacy tests; 3\*\*\**p*<0.001; \*\**p*<0.005; \**p*<0.05 4No percentile growth scores are available for STE.

The magnitude of the relationships between school climate scores and achievement (Table 12) and other criterions (Table 11) is similar to what has been reported previously (Peoples, 2016; Hough, Kalogrides, and Loeb, 2017; Peoples, Flanagan, and Foster, 2017).

* + 1. **External validity conclusion**

Overall, the external validity evidence supports the conclusion that the school climate surveys are responsive (at the individual-level and at the school level) and can measure change in student perceptions of school climate. Hough, Kalogrides, and Loeb (2017) found that most of the variation in students’ perceptions of school climate were within rather than between schools. The proportion of variance explained in Massachusetts between schools in grade 5, grade 8, and grade 10 was 10.0%, 9.1%, and 9.5%, respectively; these are of the same magnitude to those found by Hough, Kalogrides, and Loeb (2017). Based on their analyses of the CORE districts in California, they recommended using three levels to characterize school “performance”. Massachusetts schools were divided into three “performance” levels; this division ensures that schools can be meaningfully characterized and differentiated. The school climate measures and profiles provided to schools are intended to support a continuous improvement process.

The pattern of correlations provides preliminary evidence to support VOCAL’s external validity; replication these associations across grades and years strengthens the external validity argument. However, the correlational cross-sectional data do not support the interpretation that more positive school climates lead to (cause) improved student achievement. In addition, these simple correlations do not account for the nested nature of educational data. Future validity work will focus on providing external validity evidence using hierarchical linear models that consider the nested structure of education data and assess the VOCAL scaled scores predictive validity.

* 1. **Consequential validity**

Consequential validity discusses the implications of using the scores for their intended purpose. It “appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use” (Messick, 1995b, p.6). The Massachusetts Safe and Supportive Schools Commission (2019a, p.1) advocates that, “safe and supportive school environments are essential in order to reach high academic standards and other important educational reform goals, including diminishing the use of suspension and expulsion as an approach to discipline, preventing bullying, preventing substance use and misuse and providing support for addiction recovery, closing proficiency gaps, and halting the school to prison pipeline.” The VOCAL survey was designed to provide schools and districts with a measure of how safe and supportive their school environments are. DESE’s primary goal is for educators to use the VOCAL data for continuous school improvement; in addition, the school climate data helps DESE meet the survey requirement of section 370 of the [Act Relative to Bullying in Schools](https://malegislature.gov/Laws/GeneralLaws/PartI/TitleXII/Chapter71/Section37O).

At this time, there are no high stakes decisions or risks associated with the use of the survey scores; participation by students, schools, and districts is voluntary, and the data are not part of the state’s accountability system. Student confidentiality is protected as schools and districts only receive aggregate results and only if they meet DESE’s minimum reporting criteria of an N of 10 and a school-level or district-level person separation reliability of 0.7 or more. The consequences for individual students are minimal as student-level information is not subject to public records requests. However, with aggregate data subject to public records request and with the survey used to comply with the Act Relative to Bullying in Schools, there are potential consequences attached to the use of the scores. A discussion of the intended and some unintended consequences of the survey design and score use are discussed next.

* + 1. **Intended outcomes**

One intended outcome is for schools and districts to value the information provided and use the data to support school improvement. In 2018, a representative sample of Massachusetts educators participated in the Views of Instruction, State Standards, Teaching and Assessment (VISTA) annual survey for superintendents and principals. Educators were asked if they found the 2017 school climate reports useful; of the superintendents and principals who administered the VOCAL pilot survey, over eight in ten superintendents and principals somewhat agreed or strongly agreed that their VOCAL reports were useful ([DESE, 2018b](http://www.doe.mass.edu/research/vista/2018/voice-issue13.pdf)). Educators shared and discussed the school climate results with their staff; six in ten superintendents and almost seven in ten principals agreed that they met with staff to review their school climate results (2018b). In 2019, superintendents and principals were asked a more specific question on the VISTA survey: please evaluate the usefulness of the VOCAL school climate reports in informing your district's/school’s planning and improvement work (DESE, 2019b). Of those districts and schools who administered the VOCAL survey, almost seven in ten superintendents and three in four principals found the VOCAL data useful or very useful for school planning and improvement. Districts and schools are using the VOCAL data as intended and find the data useful for school planning and improvement.

In 2017, some schools who met the minimum N of 10 students did not receive dimension scores because they did not meet the minimum school-level reliability requirement of 0.7. This unintended consequence was ameliorated in 2018. The number of items in each dimension was increased to provide more schools with reliable school-level index scores. For example, the percent of schools who received an engagement score increased from 55% in 2017 to 77% in 2018. Of these schools who still did not receive an engagement index score in 2018, almost half (48%) did not do so because they did not meet the minimum N to report out the data. Adding items to each dimension improved the reliability of the school-level index scores. To ensure schools who have the minimum response rate receive dimension scores in 2019, the 2019 surveys were lengthened. In 2019, all students will respond to a 40-item survey that measures students’ perceptions of the three school climate dimensions.

* + 1. **Unintended outcomes**

The policy decision to use the VOCAL data to meet the survey requirement of section 370 of the [Act Relative to Bullying in Schools](https://malegislature.gov/Laws/GeneralLaws/PartI/TitleXII/Chapter71/Section37O) was not without consequence. Inclusion of several behaviorally related bullying items in the VOCAL survey led to unintended psychometric issues. These bullying behavior items (all reversed-scored) weakened the claim that the school climate items fitted the Rasch model well as they introduced error and misfit in the modeling process. These items exhibit higher than expected misfit. Additional analyses were performed to justify keeping these items in the survey; the results of these analyses (Section 5.1, p.18) indicated that, at the school-level, the overall VOCAL scores, safety, and bullying scores reported were not biased due to the retention of these items. The practical significance of including these items was minimal as the impact on reported outcomes (index scores) was negligible.

DESE did not translate the surveys into languages other than English. This decision had a negative impact on English learner students and unintended psychometric issues. English learners were not able to access some items as their grammatical structure and language was too complex. English learner scores were based on fewer items than other subgroup scores; as a result, some schools may not have received EL student scores because the reliability of these scores was lower.

The VOCAL survey was initially administered in three grades: grade 5, grade 8, and grade 10. Anecdotal feedback from educators highlighted another unintended consequence of the VOCAL survey. Massachusetts has a diversity of school configurations with some districts having elementary schools that serve students from kindergarten through grade 4; others serving students, kindergarten through grade 5, with others serving students in kindergarten through grade 8 or grade 12. Hence, districts whose elementary schools serve kindergarten through grade 4 were excluded from receiving VOCAL data. In 2019, a grade 4 survey was offered to schools.

* + 1. **Consequential validity conclusion**

The purpose of the VOCAL school climate survey is to support schools in continuously improving the school environment for their students. Educators largely agree that the VOCAL survey is serving this purpose (DESE, 2018b; DESE, 2019b). DESE has made progress in making the survey accessible to more students by the addition of grade 4. However, the lack of translating the survey into other languages undoubtedly led to English learner students not having access to the full survey. Although there were enough items that were accessible to English learners to compute their scores with reasonable accuracy, DESE should consider translating the survey into other languages in order to fully understand English learners’ school climate perceptions. In addition, DESE should consider rewording some of the reverse-scored items so they have a positive valence; this should help with the fit and structural validity of the school climate items, and may help English learners access the survey.

**6.0 VOCAL report conclusion**

The purpose of this validity study was to provide psychometric evidence to justify the use of VOCAL scores by schools and districts within Massachusetts. The conceptual framework for the VOCAL was derived from the USED’s school climate survey, a previously validated instrument. Items were included that measured the three dimensions of school climate: engagement, safety, and environment. Evidence was provided that supported each aspect of construct validity (content, structural, substantive, generalizability, external, and consequential) for the school climate measure. A large majority of the 76-item VOCAL survey fit the Rasch model well; a “scoring method” factor made up of 14 reverse-scored items misfit the model. Despite these misfitting items, the VOCAL scale met the unidimensional assumption of the Rasch model, and the presence of these items did not bias school-level scores (especially within the safety dimension where reverse-scored items were omnipresent). Students’ dimension scores (engagement, safety, and environment) were moderately to strongly correlated with each other indicating that they were distinct conceptually but structurally related by the overarching school climate construct. The rating scale structure was used by students as intended by the developers and item difficulty hierarchies for each of the dimensions met developers’ *a priori* expectations.

The VOCAL survey was reliable at the student, grade, and school level. Differential item function analyses indicated that students from different subgroups – with the exception of English learners – with the same score had, within measurement error, the same affirmation level and likely interpretation of most items. These data support the generalizability of the school climate construct. VOCAL scores were, as expected, appropriately related (positively or negatively) to other school-level non-academic criteria and positively related to students’ school level achievement. The VOCAL and dimension measures were responsive at both the student and school level. Schools were meaningfully differentiated by their school climate scores and the characterization (profile) of their “performance” was designed to give schools the information they needed to inform and support continuous improvement.

The replication of each of the validity analyses across grades and years provides further evidence that the VOCAL survey is providing reliable reproducible scores. In conclusion, the psychometric properties of the VOCAL instrument met the assumptions of the Rasch-model, namely the items are well-fitting, invariant, and form a unidimensional scale.

**References**

Andrich, D. (1978a). Application of psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2* (4), 581 –594.

Andrich, D. (1978b). Rating formulation for ordered response categories. *Psychometrika*, *43* (4), 561 –573.

Boone, W. J., and Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*, 253 –269.

Boone, W. J., Townsend, J. S., and Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data*. Science Education*, *95*, 258 –280.

Boone, W. J., Staver, J. R., and Yale, M. S. (2014). *Rasch analysis in the human sciences*, New York: Springer.

Berkowitz, R., Moore, H., Astor, R.A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate an academic achievement. *Review of Educational Research, 87* (2), 425 – 469.

Carmen, K., & Johannes, H. (2017). Practical significance of item misfit in educational assessment. *Applied Psychological Research, 41*, 5, 388 – 400.

Conrad, K. J., Wright, B. D., McKnight, P., McFall, M., Fontana, A., Rosenbeck, R. (2004). Comparing traditional and Rasch analyses of the Mississippi PTSD scale: Revealing the limitations of reverse-scored items. *Journal of Applied Measurement, 5*, (1), 1 – 16.

Crisan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the Practical Consequences of Model Misfit in Unidimensional IRT Models, Applied Psychological Measurement, 41, 6, 439 – 455.

DESE (2018a). [2017 Views of Climate and Learning (VOCAL) Validity Study](http://www.doe.mass.edu/research/vocal/2017/validity-study.docx). Available for download at <http://www.doe.mass.edu/research/vocal/2017/>

DESE (2018b). The VOICE, Office of Planning and Research. Issue 13. Available for download at <http://www.doe.mass.edu/research/vista/2018/voice-issue13.pdf>

DESE (2019) Safe and supportive schools commission – Third annual report.

DESE (2019b) 2018 – 2019 Views of Instruction, State Standards, Teaching and Assessment (VISTA) survey findings. Available at <http://www.doe.mass.edu/research/vista/2019/>

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral and health sciences.* Routledge Taylor & Francis Group, New York, New York.

Gable, R.K., Ludlow, L.H. and Wolf, M.B. (1990). The Use of Classical and Rasch Latent Trait Models to Enhance the Validity of Affective Measures. *Educational and Psychological Measurement*, *50* (4), 869 –878.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, Fall, 38 – 47.

Hafen, C.A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B. & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth Adolescence*, *41* (3), 245 –255.

Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using student surveys of students’ social and emotional learning and school climate for accountability and continuous improvement.* Policy Analysis for California Education, downloaded from http:/edpolicyinca.org.

Jennings, P. A., & Greenberg, M T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79*, 1, 491 – 525.

Johnson, S. L., Waasdorp, T. E., Debman, K., & Bradshaw, C. P. (2013). Journal of Criminology, Article ID 780460.

Linacre J. M. (2010) When to stop removing items and persons in Rasch misfit analysis?, *Rasch Measurement Transactions, 23:4,* 1241

Linacre, J. M. (2017). *A user’s guide to Winsteps, Ministep Rasch-model computer programs: program manual* 4.0.0, Chicago, US: MESA Press.

Ludlow, L. H. & Haley, S. M. (1995). Rasch model logits: Interpretation, use and transformation. *Educational and Psychological Measurement*, *55* (6), 967 – 975.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012 – 1027.

Messick, S. (1995a). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50* (9), 741 – 749.

Messick, S. (1995b). Standards of validity and the validity of standards in performance assessment. *Educational Measurement, Issues and Practice. 14*, 5 – 8

Peoples, S. M., O’Dwyer, L. M., Wang, Y., Brown, J. & Rosca, C. V. (2014) Development and Application of the Elementary School Science Classroom Environment Scale (ESSCES): Measuring Student Perceptions of Constructivism within the Science Classroom, *Learning Environments Research Journal*, 17, (1), 49 – 73.

Peoples, S.M., Abbott, C., and Flanagan, K. (2015a). *Developing student feedback surveys for educator evaluation: Combining stakeholder engagement and psychometric analyses in their development.* Paper presented to the April, 2015 annual meeting of the American Educational Research Association, Chicago, IL, US.

Peoples, S.M., Abbott, C., and Flanagan, K. (2015b). *Developing student feedback surveys for educator evaluation: Validating student feedback surveys for educator evaluation using Rasch survey development tools and the Rasch construct validity framework.* Paper presented at the April, 2015 annual meeting of the American Educational Research Association, Chicago, IL, US.

Peoples, S. (2016). *College and Career Readiness Mathematical Practice Scale CCRMS: Assessing middle and high school students’ mathematics self-efficacy.* Paper presented at American Educational Research Association Conference, Washington, DC, 2016, District of Columbia.

Peoples, S., Flanagan, K., & Foster, B. (2017). *Measuring students’ college and career readiness in English Language Arts using a Rasch-based self-efficacy scale.* Paper presented at American Educational Research Association Conference, San Antonio, Texas, 2017.

Polanin, J. R., Espelage, D. L., Pigott, T. D. (2012). A Meta-analysis of school-based bullying prevention programs’ effects on bystander intervention behavior. *School Psychology Review, 41*, 1, 47 65.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press).

Smith, E. V. Jr. (2000). Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement*, *1*(3), 303 – 326.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement,* *3*, 205 – 231.

Smith, A. B. Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*, 33 – 44

Schumacker, R. E. & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement*, *67* (3), 394 – 409.

Schwartz, R., Ayers, E., & Wilson, M. (2017). Mapping a data modeling and statistical reasoning learning progression using unidimensional and multidimensional item response models. *Journal of Applied Measurement, 18(3),* 268 – 298.

Sinharay, S., and Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice. 33*, 23 – 35.

Sinnema, C. E. L. and Ludlow, L. H. (2013). A Rasch approach to the measurement of responsive curriculum practice in the context of curricula reform. *The International Journal of Educational and Psychological Assessment, 12* (2), 33 – 55.

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.

Thomas, G. P. (2004). Dimensionality and construct validity of an instrument designed to measure the metacognitive orientation of science classroom learning environments. *Journal of Applied Measurement*, *5*(4), 367 – 384.

TNTP (2018). *The opportunity myth*. New York, NY: Author. Retrieved July 20, 2019, from https://tntp.org/assets/documents/TNTP\_The-Opportunity-Myth\_Web.pdf

United States Department of Education. (2019). National Center on Safe Supportive Learning Environments, ED School Climate Surveys (EDSCLS), <https://safesupportivelearning.ed.gov/edscls/measures>

Van Rijn, P. W., Sinharay, S., Haberman, S. J. and Johnson, M. S. (2016). Assessment of fit of item response theory models used in large‑scale educational survey assessments. *Large Scale Assessments in Education, 4,* (10), 1 – 23.

Wolfe, E. W., & Smith, E. V. Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – Instrument development tools. *Journal of Applied Measurement*, *8* (1), 97 – 123.

Wolfe, E. W. & Smith Jr., E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. *Journal of Applied Measurement*, *8* (2), 204 – 234.

Wright, B.D., and Stone, M. H. (1979). Best test design: Rasch measurement. MESA press, Chicago, Illinois.

Wright B.D., & Linacre J.M. (1994) Reasonable mean-square fit values. *Rasch Measurement Transactions, 8:3* p.370

Wright, B. D., and Masters, G. N. (2002). Number of Person or Item Strata. *Rasch Measurement Transactions*, *16* (3), 888.

**APPENDICES**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Appendix A: VOCAL 2018 survey specification (common items are only counted once)** | | | | | |
|  |  | **G5 items** | **G8 items** | **G10 items** | **Total** |
| **Dimension** | **Domains** | **Number** | **Number** | **Number** | **Number** |
| **Engagement (ENG)** | **Cultural and linguistic competence (CLC)** | **3** | **4** | **4** | **7** |
| **Engagement (ENG)** | **Relationships (REL)** | **3** | **4** | **4** | **6** |
| **Engagement (ENG)** | **Class and school participation (PAR)** | **6** | **4** | **4** | **12** |
| **Dimension** | **Domains** | **12** | **12** | **12** | **25** |
| **Safety (SAF)** | **Emotional safety (EMO)** | **4** | **5** | **5** | **10** |
| **Safety (SAF)** | **Physical safety (PSF)** | **2** | **2** | **2** | **6** |
| **Safety (SAF)** | **Bullying/cyber-bullying (BUL)** | **7** | **8** | **8** | **13** |
| **Dimension** | **Domains** | **13** | **15** | **15** | **29** |
| **Environment (ENV)** | **Instructional environment (INS)** | **6** | **6** | **6** | **10** |
| **Environment (ENV)** | **Mental health environment (MEN)** | **2** | **2** | **2** | **6** |
| **Environment (ENV)** | **Discipline environment (DIS)** | **3** | **3** | **3** | **5** |
| **Subtotal** |  | **11** | **11** | **11** | **22** |
| **TOTAL** |  | **36** | **38** | **38** | **76** |

**Appendix B1: Student MCAS questionnaire - Grade 5 VOCAL form items**

Table includes how each item was scored; items are reverse-scored when greater affirmation of the item by the student indicates a more negative school climate. Items highlighted in green are common across all three grade-level forms.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Think of the last 30 days in school.** | | **Always true** | **Mostly true** | **Mostly untrue** | **Never true** |
| 1. | | Teachers support (help) students who come to class upset. | | 3 | 2 | 1 | 0 |
| 2. | | My school work is challenging (hard) but not too difficult. | | 3 | 2 | 1 | 0 |
| 3. | | I feel safe at our school. | | 3 | 2 | 1 | 0 |
| 4. | | When I am stuck, my teachers want me to try again before they help me. | | 3 | 2 | 1 | 0 |
| 5. | | My teachers care about me as a person. | | 3 | 2 | 1 | 0 |
|  | **Think of the last 30 days in school.** | | **Always true** | | **Mostly true** | **Mostly untrue** | **Never true** |
| 6. | Teachers give students a chance to explain their behavior when they do something wrong. | | 3 | | 2 | 1 | 0 |
| 7. | In the last month, I have seen **more than one** physical fight at my school. | | 0 | | 1 | 2 | 3 |
| 8. | Students respect one another. | | 3 | | 2 | 1 | 0 |
| 9. | Teachers **don’t** let students pick on other students in class or in the hallways. | | 3 | | 2 | 1 | 0 |
| 10. | My teachers are proud of me when I work hard in school. | | 3 | | 2 | 1 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | **Mostly untrue** | **Never true** |
| 11. | In my school, groups of students tease or pick on one student. | 0 | 1 | 2 | 3 |
| 12. | I get the chance to take part in school events (for example, science fairs, art or music shows). | 3 | 2 | 1 | 0 |
| 13. | School rules are fair for all students. | 3 | 2 | 1 | 0 |
| 14. | Adults working at this school treat **all** students respectfully. | 3 | 2 | 1 | 0 |
| 15. | Students help each other learn without having to be asked by the teacher. | 3 | 2 | 1 | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | **Mostly untrue** | **Never true** |
| 16. | My teachers will explain things in different ways until I understand. | 3 | 2 | 1 | 0 |
| 17. | If I tell a teacher or other adult at school that someone is being bullied, the teacher/adult will do something to help. | 3 | 2 | 1 | 0 |
| 18. | I am happy to be at our school. | 3 | 2 | 1 | 0 |
| 19. | Students have a voice in deciding school rules. | 3 | 2 | 1 | 0 |
| 20. | Students will help other students if they are upset, even if they are not close friends. | 3 | 2 | 1 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | **Mostly untrue** | **Never  true** |
| 21. | My teachers use my ideas to help my classmates learn. | 3 | 2 | 1 | 0 |
| 22. | At our school, students learn to care about other students' feelings. | 3 | 2 | 1 | 0 |
| 23. | My teachers ask me to share what I have learned in a lesson. | 3 | 2 | 1 | 0 |
| 24. | Teachers, students, and the principal work together in our school to prevent (stop) bullying. | 3 | 2 | 1 | 0 |
| 25. | Teachers at this school accept me for who I am. | 3 | 2 | 1 | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | **Mostly untrue** | **Never  true** |
| 26. | I feel comfortable talking to my teacher(s) about something that is bothering me. | 3 | 2 | 1 | 0 |
| 27. | In school, I learn how to manage (control) my feelings when I am angry or upset. | 3 | 2 | 1 | 0 |
| 28. | When I need help, my teachers use my interests to help me learn. | 3 | 2 | 1 | 0 |
| 29. | Students at school try to stop bullying when they see it happening. | 3 | 2 | 1 | 0 |
| 30. | My teachers support me even when my work is not my best. | 3 | 2 | 1 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | **Mostly untrue** | **Never  true** |
| 31. | In my school, older students scare or pick on younger students. | 0 | 1 | 2 | 3 |
| 32. | When I am home, I like to learn more about the things we are learning in school. | 3 | 2 | 1 | 0 |
| 33. | Students like to have friends who are different from themselves (for example, boys and girls, rich and poor, or classmates of different color). | 3 | 2 | 1 | 0 |
| 34. | I have been punched or shoved by other students **more than once** in the school or on the playground. | 0 | 1 | 2 | 3 |
| 35. | Students at my school get along well with each other. | 3 | 2 | 1 | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | **Mostly untrue** | **Never  true** |
| 36. | My teachers help me succeed with my school work when I need help. | 3 | 2 | 1 | 0 |

**Appendix B2: Student MCAS questionnaire - Grade 8 VOCAL form items**

Table includes how each item was scored; items are reverse-scored when greater affirmation of the item by the student indicates a more negative school climate Items highlighted in green are common across all three grade-level forms.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | | **Mostly untrue** | **Never true** | | |
| 1. | Teachers support (help) students who come to class upset. | 3 | 2 | | 1 | 0 | | |
| 2. | My school work is challenging (hard) but not too difficult. | 3 | 2 | | 1 | 0 | | |
| 3. | I have a choice in how I show my learning (e.g., write a paper, prepare a presentation, make a video). | 3 | 2 | | 1 | 0 | | |
| 4. | My teachers believe that **all** students can do well in their learning. | 3 | 2 | | 1 | 0 | | |
| 5. | Teachers are available when I need to talk with them. | 3 | 2 | | 1 | 0 | | |
| 6. | Teachers give students a chance to explain their behavior when they do something wrong. | 3 | 2 | | 1 | 0 | | |
| 7. | Students have spread rumors or lies about me **more than once** on social media. | 0 | 1 | | 2 | 3 | | |
| 8. | Students respect one another. | 3 | 2 | | 1 | 0 | | |
| 9. | Teachers **don’t** let students pick on other students in class or in the hallways. | 3 | 2 | | 1 | 0 | | |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | **Mostly untrue** | | **Never true** |
| 10. | My teachers are proud of me when I work hard in school. | 3 | | 2 | 1 | | 0 |
| 11. | In my school, groups of students tease or pick on one student. | 0 | | 1 | 2 | | 3 |
| 12. | In my class, my teacher uses students' interests to plan class activities. | 3 | | 2 | 1 | | 0 |
| 13. | If I need help with my emotions (feelings), **effective** help is available at my school. | 3 | | 2 | 1 | | 0 |
| 14. | Adults working at this school treat **all** students respectfully, regardless of a student's race, culture, family income, religion, sex, or sexual preference. | 3 | | 2 | 1 | | 0 |
| 15. | Students help each other learn without having to be asked by the teacher. | 3 | | 2 | 1 | | 0 |
| 16. | Because I worry about my grades, it is hard for me to enjoy school. | 0 | | 1 | 2 | | 3 |
| 17. | If I tell a teacher or other adult at school that someone is being bullied, the teacher/adult will do something to help. | 3 | | 2 | 1 | | 0 |
| 18. | My textbooks or class materials include people and examples that reflect my race, cultural background and/or identity. | 3 | | 2 | 1 | | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | **Mostly untrue** | | **Never true** |
| 19. | Students have a voice in deciding school rules. | 3 | | 2 | 1 | | 0 |
| 20. | Students will help other students if they are upset, even if they are not close friends. | 3 | | 2 | 1 | | 0 |
| 21. | My teachers use my ideas to help my classmates learn. | 3 | | 2 | 1 | | 0 |
| 22. | My teachers set high expectations for my work. | 3 | | 2 | 1 | | 0 |
| 23. | Students at school damage and/or steal other students' property. | 0 | | 1 | 2 | | 3 |
| 24. | Teachers, students, and the principal work together in our school to prevent (stop) bullying. | 3 | | 2 | 1 | | 0 |
| 25. | My teachers promote respect among students. | 3 | | 2 | 1 | | 0 |
| 26. | In my school, bigger students taunt or pick on smaller students. | 0 | | 1 | 2 | | 3 |
| 27. | I feel comfortable reaching out to teachers/counselors for emotional support if I need it. | 3 | | 2 | 1 | | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | **Mostly untrue** | | **Never  true** |
| 28. | Students from different backgrounds respect each other in our school, regardless of their race, culture, family income, religion, sex, or sexual preference. | 3 | | 2 | 1 | | 0 |
| 29. | Students at school try to stop bullying when they see it happening. | 3 | | 2 | 1 | | 0 |
| 30. | My teachers support me even when my work is not my best. | 3 | | 2 | 1 | | 0 |
| 31. | Our school offers guidance to students on how to mediate (settle) conflicts (e.g., arguments, fights) by themselves. | 3 | | 2 | 1 | | 0 |
| 32. | Teachers and adults are interested in my well-being beyond just my class work. | 3 | | 2 | 1 | | 0 |
| 33. | Students are open to having friends who come from different backgrounds (for example, friends from different races, cultures, family incomes, or religions, or friends of a different sex or sexual preference). | 3 | | 2 | 1 | | 0 |
| 34. | Adults at our school are respectful to student ideas even if the ideas expressed are different from their own. | 3 | | 2 | 1 | | 0 |
| 35. | I have seen students with weapons at our school. | 0 | | 1 | 2 | | 3 |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | **Mostly untrue** | | **Never  true** |
| 36. | I have been called names or made fun of by other students **more than once** in school. | 0 | | 1 | 2 | | 3 |
| 37. | My parents feel respected when they participate at our school (e.g., at parent-teacher conferences, open houses). | 3 | | 2 | 1 | | 0 |
| 38. | School staff are consistent when enforcing rules in school. | 3 | | 2 | 1 | | 0 |

**Appendix B3: Student MCAS questionnaire - Grade 10 VOCAL form**

Table includes how each item was scored; items are reverse-scored when greater affirmation of the item by the student indicates a more negative school climate. Items highlighted in green are common across all three grade-level forms.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Think of the last 30 days in school.** | **Always true** | | | **Mostly true** | | | **Mostly untrue** | | **Never true** | | |
| 1. | Teachers support (help) students who come to class upset. | 3 | | | 2 | | | 1 | | 0 | | |
| 2. | I feel as though I belong in my school community. | 3 | | | 2 | | | 1 | | 0 | | |
| 3. | My teachers inspire confidence in my ability to be ready for college or career. | 3 | | | 2 | | | 1 | | 0 | | |
| 4. | In at least two of my academic classes, I can work on assignments that interest me personally. | 3 | | | 2 | | | 1 | | 0 | | |
| 5. | Teachers are available when I need to talk with them. | 3 | | | 2 | | | 1 | | 0 | | |
| 6. | Teachers give students a chance to explain their behavior when they do something wrong. | 3 | | | 2 | | | 1 | | 0 | | |
| 7. | I feel welcome to participate in extra-curricular activities offered through my school, such as school clubs or organizations, musical groups, sports teams, student council, or any other extra-curricular activities. | 3 | | | 2 | | | 1 | | 0 | | |
| 8. | Students respect one another. | 3 | | | 2 | | | 1 | | 0 | | |
|  | **Think of the last 30 days in school.** | **Always true** | **Mostly true** | | | **Mostly untrue** | | | **Never true** | | |
| 9. | Teachers **don’t** let students pick on other students in class or in the hallways. | 3 | | 2 | | | 1 | | | | 0 |
| 10. | The consequences for the **same** inappropriate behavior (e.g., disrupting the class) are the same, no matter who the student is. | 3 | | 2 | | | 1 | | | | 0 |
| 11. | In my school, groups of students tease or pick on one student. | 0 | | 1 | | | 2 | | | | 3 |
| 12. | I have access to **effective** help at school if I am struggling emotionally or mentally. | 3 | | 2 | | | 1 | | | | 0 |
| 13. | I have a group of friends I can rely on to help me when I feel down (sad). | 3 | | 2 | | | 1 | | | | 0 |
| 14. | Adults working at this school treat **all** students respectfully, regardless of a student’s race, culture, family income, religion, sex, or sexual preference. | 3 | | 2 | | | 1 | | | | 0 |
| 15. | Students help each other learn without having to be asked by the teacher. | 3 | | 2 | | | 1 | | | | 0 |
| 16. | Because I worry about my grades, it is hard for me to enjoy school. | 0 | | 1 | | | 2 | | | | 3 |
| 17. | If I tell a teacher or other adult at school that someone is being bullied, the teacher/adult will do something to help. | 3 | | 2 | | | 1 | | | | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | | | **Mostly untrue** | | | | **Never  true** |
| 18. | Students are sexually harassed at my school (for example, bothered by unwanted touching and/or indecent name-calling). | 0 | | 1 | | | 2 | | | | 3 |
| 19. | Students have a voice in deciding school rules. | 3 | | 2 | | | 1 | | | | 0 |
| 20. | I am encouraged to take upper level courses (honors, AP). | 3 | | 2 | | | 1 | | | | 0 |
| 21. | My teachers use my ideas to help my classmates learn. | 3 | | 2 | | | 1 | | | | 0 |
| 22. | My teachers set high expectations for my work. | 3 | | 2 | | | 1 | | | | 0 |
| 23. | I have stayed at home (or avoided school) because I did not feel safe at my school. | 0 | | 1 | | | 2 | | | | 3 |
| 24. | Teachers, students, and the principal work together in our school to prevent (stop) bullying. | 3 | | 2 | | | 1 | | | | 0 |
| 25. | My teachers promote respect among students. | 3 | | 2 | | | 1 | | | | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | | | **Mostly untrue** | | | | **Never  true** |
| 26. | I have been teased or picked on **more than once** because of my real or perceived (imagined) sexual preference. | 0 | | 1 | | | 2 | | | | 3 |
| 27. | The level of pressure I feel at school to perform well is unhealthy. | 0 | | 1 | | | 2 | | | | 3 |
| 28. | Students from different backgrounds respect each other in our school, regardless of their race, culture, family income, religion, sex, or sexual preference. | 3 | | 2 | | | 1 | | | | 0 |
| 29. | Students at school try to stop bullying when they see it happening. | 3 | | 2 | | | 1 | | | | 0 |
| 30. | My teachers support me even when my work is not my best. | 3 | | 2 | | | 1 | | | | 0 |
| 31. | I have been teased or picked on **more than once** because of my race or ethnicity. | 0 | | 1 | | | 2 | | | | 3 |
| 32. | Teachers ask students for feedback on their classroom instruction. | 3 | | 2 | | | 1 | | | | 0 |
| 33. | Students are open to having friends who come from different backgrounds (for example, friends from different races, cultures, family incomes, or religions, or friends of a different sex, or sexual preference). | 3 | | 2 | | | 1 | | | | 0 |
|  | **Think of the last 30 days in school.** | **Always true** | | **Mostly true** | | | **Mostly untrue** | | | | **Never  true** |
| 34. | Adults at our school are respectful to student ideas even if the ideas expressed are different from their own. | 3 | | 2 | | | 1 | | | | 0 |
| 35. | If I finish my work early, I have an opportunity to do more challenging work. | 3 | | 2 | | | 1 | | | | 0 |
| 36. | The things I am learning in school are relevant (important) to me. | 3 | | 2 | | | 1 | | | | 0 |
| 37. | Students with learning or physical difficulties are teased or picked on at my school. | 0 | | 1 | | | 2 | | | | 3 |
| 38. | Students at school try to work out their problems with other students in a respectful way. | 3 | | 2 | | | 1 | | | | 0 |

**Appendix C1: The Rasch model**

The Rasch model uses an exponential transformation to place ordinal Likert responses on to an equal-interval logit scale (Rasch, 1960). This transformation ensures that stakeholder perceptions are measured appropriately, and that the data meet the assumptions of parametric testing (Ludlow and Haley, 1995; Boone, Staver, and Yale 2014). In addition, the sample independence features of the Rasch model overcome the fundamental drawbacks of classical test theory (CTT) analyses Smith (2000). In CTT, the difficulty of a test is sample dependent, making it problematic to measure change on a variable (Smith, 2000; Boone & Scantlebury, 2006). In contrast, the Rasch property of item invariance implies that the relative endorsements and location of the items do not change (within measurement error) or are independent of the sample responding; in kind, the relative item endorsements should behave as expected across different samples (Smith, 2002, Engelhard, 2013). When items are invariant, the Rasch model is particularly discerning in differentiating between high and low scorers (Gable, Ludlow, and Wolf, 1990; Sinnema & Ludlow, 2013) on a measurement scale as it places persons and items on a common scale metric (Hambleton and Jones, 1993; Engelhard, 2013).

The Rasch rating scale model provides a mathematical model for the probabilistic relationship between a person’s ability () and the difficulty of items () on a test or survey. Andrich’s (1978a, 1978b) rating scale model (RSM) used in this study is defined in Equation 1.

j = 1, 2, …, mi. (1)

Where is the “conditional probability of person, *n* responding in category *j* to item *i*”. Tau is the estimate of the location of the *jth* step for each item relative to that item’s scale value (*δi*). The number of response categories is equal to *mi* +1 where *mi* is the number of thresholds. In the RSM, moving from one threshold to the next contiguous threshold is assumed to have the same mean difference across all items of the survey. The unit of measurement resulting from the natural log transformation of person responses results in separate ability and item difficulty estimates called logits (Ludlow & Haley, 1995). The persons and items are placed on a common continuum (the scale metric axis of the variable map) and as such, the persons can be characterized by their location on the continuum by the types and level of items of which they are associated. By taking the natural log of the odds ratio, stable replicable information about the relative strengths of persons and items is derived with equal differences in logits translating into equal differences in the probability of endorsing an item no matter where on the scale metric an item is located; this interval-level unit of measurement is a fundamental assumption of parametric tests (Boone, Townsend, and Staver, 2011). By default, in WINSTEPS, the item means summed across the thresholds equals zero; the person and item measures are generated and reported on the logit scale. In the context of this study, a respondent with a positive logit value on an educator preparation survey feels relatively more positive about the program than a respondent with a negative logit value.

Appendix C2: Logit unit of measurement

The unit of measurement resulting from the natural log transformation of person responses results in separate ability and item difficulty estimates called logits (Ludlow & Haley, 1995); this transformation expands the theoretical ability (endorsement) range from negative infinity to plus infinity with most estimates falling in the range of -4 to +4 logits (Ludlow & Haley, 1995). Items can be similarly interpreted in logits with a theoretical range of negative infinity to positive infinity; items with a positive logit are, on average, more difficult to endorse than items with negative logits (Ludlow & Haley, 1995). The persons and items are placed on a common continuum (the scale metric axis of the variable map) and as such, the persons can be characterized by their location on the continuum by the types and level of items of which they are associated. Person expected responses can be compared to their observed responses to determine if “the logit estimate of ability (affirmation) corresponding to an original raw data summary score is consistent or inconsistent with the pattern expected for that estimate of ability (affirmation)” (Ludlow & Haley, 1995). By taking the natural log of the odds ratio, stable replicable information about the relative strengths of persons and items is derived with equal differences in logits translating into equal differences in the probability of endorsing an item no matter where on the scale metric an item is located; this interval-level unit of measurement is a fundamental assumption of parametric tests (Ludlow and Haley, 1995; Boone, Townsend, and Staver, 2011).

**Appendix D: Guide for evaluating Rasch model validity data**

| **Validity Aspect** | **Statistic/Data** | **Cutoff Criteria or Typical Standard** | **Comment** |
| --- | --- | --- | --- |
| **Content** | Point-to-measure Correlation | Positive and >0.3. | Analog to CTT item-total correlation. |
| **Content & Structural** | Outfit mean-square fit statistic (MNSQ)  Linacre, 2019 | * 0.5 – 1.5 productive for measurement * 1.5 – 2.0 unproductive for construct, but does not degrade measurement * >2.0 distorts or degrades measure * <0.5 not as productive for construct but does not distort measures. | Mean square errors should have a mean of one i.e. (observed = expected).  Mean square is a chi-square statistic adjusted for sample size. |
| **Substantive** | Rating scale functioning | * Minimum of 10 responses per category. * Categories are unimodal. * Observed score averages and item threshold parameters increase monotonically. * Un-weighted MNSQ < 2.0 for ea. category. | Rating scale is used according to the intent of instrument developers – supports score use and inferences. |
| **Substantive** | Item difficulty hierarchy | * Ordering of item deltas correspond to theoretical expectations. * Item/person variable maps. | Qualitative assessment of items in the construct and/or dimensions/domains. |
| **Generaliz-ability** | Item invariance and  Differential Item Functioning (DIF) | * Within standard error, items should retain same item difficulty (deltas) across administrations and survey forms (correlation of greater than or equal to 0.9). * For DIF, recommended criteria vary: delta difference of 0.3 – 0.67 logits (**0.5 used in study**) | DIF flags items that need further review. Items may need revision to eliminate bias or removal when estimating scores if bias is significant. |
| **Generaliz-ability** | Person separation  reliability (PSR) | * Typical ~ 0.8; High Stakes > 0.9 * 0.9 Construct; 0.8 Dimensions; 0.7 school-level scores | PSR is similar to Cronbach α and ranges from 0 to 1. |
| **Structural** | Sub-scale correlations | * Positive and substantial (> 0.5 but < 0.9) |  |
| **Structural** | Standardized Residuals | * No correlation between residuals from separate calibrations of two item subsets. |  |
| **Structural** | Winsteps Software  (PCA: Principal component analyses of residuals). | * Total variance explained:   >40% very good; >50% excellent   * 2nd dimension: < 5% of total variance. * 2nd dimension Eigen < 3 * 1st contrast item variance 4x variance of 2nd item contrast * Cluster correlations   + - > 0.82 likely only one latent trait   + > 0.71 more dependency than independence | The items that form a 2nd dimension should be reviewed qualitatively to determine their commonality and if their co-variation is meaningful. |
| **External** | Responsiveness | * Typical ~ 3 person strata (low, medium, high). * H = (4G +1)/3 where H is the number of person strata and G is the person separation index. | Instruments that are responsive can better differentiate high and low scorers by reliably separating individuals into a greater number of performance levels, thereby facilitating the measurement of change of respondent views on a construct. |

**Appendix E1: Technical quality (mean-square error) of 76-item VOCAL scale**

**-------------------------------------------------------------------------------------------------------**

**|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |**

**|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |**

**|------------------------------------+----------+----------+-----------+-----------+--------+---------|**

**| 40 124378 59199 .35 .01|2.23 9.90|2.24 9.90|A .36 .54| 33.7 54.5| .00| SAFPSF7 |**

**| 46 137333 62705 -.44 .01|2.20 9.90|2.19 9.90|B .33 .48| 34.7 55.4| .00| SAFBUL5 |**

**| 48 111398 42839 -1.65 .01|2.19 9.90|2.10 9.90|C .29 .39| 64.8 66.6| .00| SAFBUL10|**

**| 49 107771 42852 -1.35 .01|2.17 9.90|2.16 9.90|D .28 .42| 54.9 62.4| .00| SAFBUL11|**

**| 43 122153 59169 .43 .01|2.16 9.90|2.15 9.90|E .42 .55| 32.6 54.1| .00| SAFBUL2 |**

**| 39 107212 42912 -1.30 .01|1.90 9.90|1.83 9.90|F .37 .42| 57.3 61.8| .00| SAFPSF5 |**

**| 50 129234 59174 .16 .01|1.84 9.90|1.89 9.90|G .34 .53| 44.5 55.5| .00| SAFBUL12|**

**| 38 156471 62604 -1.28 .01|1.85 9.90|1.81 9.90|H .35 .43| 56.6 61.7| .00| SAFPSF4 |**

**| 3 103805 62113 .63 .01|1.70 9.90|1.81 9.90|I .25 .53| 39.5 49.5| .00| ENGCLC3 |**

**| 52 112656 62622 .39 .01|1.80 9.90|1.80 9.90|J .44 .52| 33.7 51.2| .00| SAFBUL14|**

**| 35 126168 105625 1.48A .00|1.49 9.90|1.68 9.90|K .33 .55| 39.2 45.8| -.01| SAFEMO11|**

**| 4 87939 42837 -.14 .01|1.64 9.90|1.67 9.90|L .40 .50| 40.9 54.4| .00| ENGCLC4 |**

**| 8 140245 59186 -.32 .01|1.67 9.90|1.64 9.90|M .38 .50| 51.6 59.2| .00| ENGPAR1 |**

**| 71 67141 42955 .81 .01|1.53 9.90|1.58 9.90|N .43 .54| 38.7 48.2| .00| ENVMEN9 |**

**| 30 105128 42832 -1.16 .01|1.57 9.90|1.48 9.90|O .39 .43| 55.7 60.5| .00| SAFEMO6 |**

**| 54 91346 42719 -.33 .01|1.48 9.90|1.53 9.90|P .38 .49| 47.7 55.1| .00| SAFBUL16|**

**| 41 93477 42820 -.44 .01|1.38 9.90|1.41 9.90|Q .38 .48| 51.9 55.8| .00| SAFPSF8 |**

**| 28 97433 62610 .84A .01|1.39 9.90|1.40 9.90|R .53 .54| 39.4 48.2| .00| SAFEMO4 |**

**| 51 294465 164814 .61A .00|1.33 9.90|1.39 9.90|S .41 .56| 46.9 51.0| .00| SAFBUL13|**

**| 72 179258 164702 1.90A .00|1.33 9.90|1.39 9.90|T .44 .58| 44.3 47.2| -.01| ENVDIS1 |**

**| 16 141117 59201 -.37A .01|1.36 9.90|1.38 9.90|U .31 .50| 51.8 59.4| .01| ENGPAR9 |**

**| 64 85682 59176 1.64 .01|1.30 9.90|1.35 9.90|V .48 .58| 44.1 48.4| .00| ENVINS14|**

**| 6 141284 59158 -.38A .01|1.30 9.90|1.29 9.90|W .42 .50| 57.2 59.9| .01| ENGCLC6 |**

**| 32 114929 59168 .70A .01|1.29 9.90|1.26 9.90|X .56 .56| 47.2 52.7| .00| SAFEMO8 |**

**| 53 126323 62642 -.05 .01|1.28 9.90|1.29 9.90|Y .40 .50| 52.0 53.7| .00| SAFBUL15|**

**| 10 103150 42980 -1.00 .01|1.27 9.90|1.21 9.90|Z .47 .44| 57.3 59.0| .00| ENGPAR3 |**

**| 66 134770 59171 -.08A .01|1.27 9.90|1.20 9.90| .52 .52| 55.2 57.2| .01| ENVMEN1 |**

**| 60 233500 122032 .44A .00|1.13 9.90|1.26 9.90| .21 .55| 54.2 52.4| .00| ENVINS9 |**

**| 9 140206 62276 -.59A .01|1.24 9.90|1.18 9.90| .51 .48| 56.1 56.0| .01| ENGPAR2 |**

**| 18 82115 42980 .15A .01|1.23 9.90|1.23 9.90| .46 .51| 45.0 52.7| .00| ENGPAR11|**

**| 68 125647 62661 -.03A .01|1.22 9.90|1.18 9.90| .57 .51| 48.1 53.4| .00| ENVMEN4 |**

**| 14 117259 59168 .61A .01|1.20 9.90|1.17 9.90| .55 .56| 48.7 53.2| .00| ENGPAR7 |**

**| 36 150540 59204 -.86A .01|1.20 9.90|1.09 9.90| .51 .47| 67.2 65.5| .00| SAFPSF1 |**

**| 67 129398 62551 -.16A .01|1.18 9.90|1.16 9.90| .52 .50| 51.0 54.1| .00| ENVMEN3 |**

**| 73 139648 59195 -.30A .01|1.18 9.90|1.13 9.90| .55 .51| 59.6 58.8| .01| ENVDIS2 |**

**| 15 116919 59177 .62A .01|1.16 9.90|1.14 9.90| .46 .56| 51.9 53.2| .01| ENGPAR8 |**

**| 37 111186 62666 .44 .01|1.12 9.90|1.15 9.90| .44 .52| 51.6 50.9| .00| SAFPSF3 |**

**| 19 70987 42821 .64A .01|1.11 9.90|1.14 9.90| .43 .54| 47.1 49.2| .00| ENGPAR12|**

**Fifteen well-fitting items removed**

**| 57 155001 59167 -1.15A .01| .98 -3.57| .82 -9.90|w .58 .44| 75.3 69.5| .01| ENVINS3 |**

**| 59 158364 62806 -1.36A .01| .98 -3.61| .90 -9.90|v .53 .43| 67.3 62.9| .01| ENVINS8 |**

**| 76 293549 164928 .63A .00| .98 -5.06| .98 -7.36|u .59 .56| 51.6 50.8| .00| ENVDIS7 |**

**| 47 379812 164703 -.54A .00| .97 -9.15| .90 -9.90|t .64 .50| 62.7 58.0| .00| SAFBUL9 |**

**| 62 67312 42773 .79A .01| .95 -8.45| .97 -5.54|s .53 .54| 50.8 48.5| .00| ENVINS12|**

**| 42 402941 164652 -.94A .00| .96 -9.60| .89 -9.90|r .59 .47| 67.2 61.4| .00| SAFBUL1 |**

**| 17 93322 62734 .97A .01| .93 -9.90| .95 -9.43|q .47 .54| 50.8 47.6| -.01| ENGPAR10|**

**| 25 251296 105544 -.94A .01| .90 -9.90| .85 -9.90|p .59 .45| 66.2 58.5| .01| ENGREL14|**

**| 61 352276 164618 -.13A .00| .90 -9.90| .88 -9.90|o .62 .52| 59.3 55.3| .00| ENVINS11|**

**| 26 346344 164995 -.04A .00| .86 -9.90| .89 -9.90|n .57 .53| 61.6 54.8| .00| SAFEMO1 |**

**| 31 92407 42979 -.36A .01| .89 -9.90| .87 -9.90|m .57 .49| 59.9 55.4| .00| SAFEMO7 |**

**| 70 127233 59174 .24A .01| .88 -9.90| .85 -9.90|l .63 .54| 60.0 55.0| .00| ENVMEN7 |**

**| 29 122626 62609 .07A .01| .87 -9.90| .86 -9.90|k .63 .51| 56.4 53.0| .00| SAFEMO5 |**

**| 5 220622 105472 -.22A .00| .86 -9.90| .85 -9.90|j .55 .50| 60.5 54.5| .00| ENGCLC5 |**

**| 45 292260 164544 .63A .00| .85 -9.90| .86 -9.90|i .60 .56| 55.8 50.8| .00| SAFBUL4 |**

**| 55 326378 164864 .22A .00| .84 -9.90| .86 -9.90|h .45 .54| 59.9 53.3| .00| ENVINS1 |**

**| 34 221590 121865 .63A .00| .85 -9.90| .85 -9.90|g .56 .56| 57.2 51.3| .00| SAFEMO10|**

**| 33 67781 42702 .77A .01| .81 -9.90| .82 -9.90|f .57 .54| 55.3 48.5| .00| SAFEMO9 |**

**| 65 90212 42981 -.25A .01| .82 -9.90| .80 -9.90|e .59 .49| 59.9 54.8| .00| ENVINS15|**

**| 23 232709 105781 -.47A .00| .80 -9.90| .80 -9.90|d .54 .48| 62.2 55.7| .00| ENGREL6 |**

**| 24 218132 105459 -.17A .00| .78 -9.90| .77 -9.90|c .62 .50| 61.8 54.2| .00| ENGREL13|**

**| 20 283306 164952 .75A .00| .66 -9.90| .67 -9.90|b .56 .56| 62.7 50.1| -.01| ENGREL1 |**

**| 22 110102 59166 .86A .01| .66 -9.90| .67 -9.90|a .57 .57| 65.5 52.0| .00| ENGREL4 |**

**|------------------------------------+----------+----------+-----------+-----------+--------+---------|**

**| MEAN 165452 80801 -.06 .01|1.23 3.9|1.21 2.5| | 54.3 55.4| .00| |**

**| P.SD 88836.3 43110 .77 .00| .38 8.7| .39 9.1| | 9.5 5.5| .00| |**

**-------------------------------------------------------------------------------------------------------**

**Appendix E2: Technical quality (mean-square error) of 25 Engagement items calibrated separately**

-------------------------------------------------------------------------------------------------------

|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |

|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| 8 140245 59186 -.37 .01|1.58 9.90|1.53 9.90|A .43 .52| 52.7 59.4| .00| ENGPAR1 |

| 3 103805 62113 .65 .01|1.50 9.90|1.55 9.90|B .39 .55| 42.7 49.8| .00| ENGCLC3 |

| 4 87939 42837 -.12 .01|1.46 9.90|1.45 9.90|C .50 .54| 45.6 54.8| .00| ENGCLC4 |

| 6 141284 59158 -.38A .01|1.22 9.90|1.17 9.90|D .46 .52| 58.9 59.5| -.04| ENGCLC6 |

| 10 103150 42980 -.99 .01|1.21 9.90|1.14 9.90|E .50 .49| 60.5 60.3| .00| ENGPAR3 |

| 16 141117 59201 -.37A .01|1.21 9.90|1.21 9.90|F .41 .52| 56.3 59.4| -.04| ENGPAR9 |

| 9 140206 62276 -.59A .01|1.17 9.90|1.11 9.90|G .55 .50| 57.5 56.3| .01| ENGPAR2 |

| 18 82115 42980 .15A .01|1.16 9.90|1.17 9.90|H .52 .56| 47.8 52.9| .04| ENGPAR11|

| 2 155583 59165 -1.19A .01|1.08 9.90| .88 -9.90|I .57 .46| 75.1 69.1| -.05| ENGCLC2 |

| 19 70987 42821 .64A .01|1.04 5.92|1.07 9.90|J .51 .58| 50.3 49.5| .04| ENGPAR12|

| 1 401337 164624 -.91A .00|1.05 9.90| .98 -3.70|K .55 .49| 63.2 61.4| -.01| ENGCLC1 |

| 21 156891 59200 -1.27A .01|1.04 6.06| .88 -9.90|L .55 .46| 75.5 70.6| -.06| ENGREL3 |

| 12 108596 62797 .52A .01|1.00 -.05|1.03 5.73|M .47 .55| 52.2 50.7| .02| ENGPAR5 |

| 14 117259 59168 .61A .01|1.03 4.56| .99 -2.19|l .63 .58| 54.4 53.6| -.03| ENGPAR7 |

| 15 116919 59177 .62A .01| .98 -3.83| .96 -6.19|k .56 .58| 56.5 53.6| -.03| ENGPAR8 |

| 7 255949 105486 -1.07A .01| .97 -5.43| .94 -9.90|j .49 .48| 63.4 60.4| .01| ENGCLC7 |

| 13 144689 59192 -.54A .01| .95 -8.42| .88 -9.90|i .56 .51| 64.8 61.3| -.05| ENGPAR6 |

| 11 261745 164675 .99A .00| .87 -9.90| .87 -9.90|h .62 .59| 54.4 49.2| .01| ENGPAR4 |

| 25 251296 105544 -.94A .01| .87 -9.90| .83 -9.90|g .60 .49| 67.2 59.0| .01| ENGREL14|

| 5 220622 105472 -.22A .00| .85 -9.90| .84 -9.90|f .57 .53| 61.7 55.1| .02| ENGCLC5 |

| 17 93322 62734 .97A .01| .82 -9.90| .84 -9.90|e .56 .56| 54.1 47.6| .02| ENGPAR10|

| 23 232709 105781 -.47A .00| .79 -9.90| .81 -9.90|d .55 .51| 63.4 56.3| .01| ENGREL6 |

| 20 283306 164952 .75A .00| .71 -9.90| .75 -9.90|c .54 .58| 62.0 50.5| .00| ENGREL1 |

| 22 110102 59166 .86A .01| .72 -9.90| .74 -9.90|b .56 .59| 64.0 52.3| -.03| ENGREL4 |

| 24 218132 105459 -.17A .00| .74 -9.90| .73 -9.90|a .64 .53| 63.5 54.6| .02| ENGREL13|

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| MEAN 165572 79046 -.11 .01|1.04 .7|1.01 -1.4| | 58.7 56.3| -.01| |

| P.SD 78355.9 37642 .72 .00| .23 8.8| .23 9.2| | 7.9 5.7| .03| |

**Appendix E3: Technical quality (mean-square error) of 29 Safety items calibrated separately**

-------------------------------------------------------------------------------------------------------

|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |

|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| 48 111398 42839 -1.74 .01|1.90 9.90|1.51 9.90|A .45 .43| 67.2 67.5| .00| SAFBUL10|

| 49 107771 42852 -1.43 .01|1.90 9.90|1.64 9.90|B .44 .46| 59.3 63.8| .00| SAFBUL11|

| 40 124378 59199 .44 .01|1.89 9.90|1.79 9.90|C .52 .60| 40.5 55.5| .00| SAFPSF7 |

| 46 137333 62705 -.48 .01|1.88 9.90|1.77 9.90|D .48 .53| 40.8 56.5| .00| SAFBUL5 |

| 43 122153 59169 .52 .01|1.84 9.90|1.72 9.90|E .55 .61| 36.9 54.5| .00| SAFBUL2 |

| 38 156471 62604 -1.33 .01|1.68 9.90|1.48 9.90|F .45 .47| 60.7 63.6| .00| SAFPSF4 |

| 39 107212 42912 -1.38 .01|1.66 9.90|1.42 9.90|G .49 .46| 61.7 63.5| .00| SAFPSF5 |

| 50 129234 59174 .24 .01|1.58 9.90|1.51 9.90|H .49 .59| 49.8 57.3| .00| SAFBUL12|

| 35 126168 105625 1.48A .00|1.45 9.90|1.56 9.90|I .41 .60| 41.0 47.2| -.01| SAFEMO11|

| 30 105128 42832 -1.24 .01|1.53 9.90|1.42 9.90|J .42 .47| 56.7 61.9| .00| SAFEMO6 |

| 28 97433 62610 .84A .01|1.50 9.90|1.50 9.90|K .49 .59| 39.8 48.5| -.01| SAFEMO4 |

| 32 114929 59168 .70A .01|1.47 9.90|1.46 9.90|L .52 .62| 46.3 54.1| .09| SAFEMO8 |

| 52 112656 62622 .37 .01|1.45 9.90|1.41 9.90|M .60 .57| 38.9 51.4| .00| SAFBUL14|

| 54 91346 42719 -.40 .01|1.22 9.90|1.17 9.90|N .54 .53| 52.4 55.6| .00| SAFBUL16|

| 27 135280 59186 -.10A .01|1.16 9.90|1.10 9.90|O .59 .57| 58.9 59.7| .08| SAFEMO3 |

| 36 150540 59204 -.86A .01|1.16 9.90|1.06 6.66|n .53 .50| 68.1 67.3| .05| SAFPSF1 |

| 41 93477 42820 -.51 .01|1.13 9.90|1.08 9.90|m .54 .52| 57.0 56.4| .00| SAFPSF8 |

| 51 294465 164814 .61A .00|1.09 9.90|1.10 9.90|l .57 .61| 52.3 51.4| .01| SAFBUL13|

| 44 374305 164873 -.45A .00|1.09 9.90|1.06 9.90|k .56 .54| 59.5 58.4| -.01| SAFBUL3 |

| 53 126323 62642 -.08 .01|1.06 9.90|1.02 3.07|j .54 .55| 55.6 54.0| .00| SAFBUL15|

| 29 122626 62609 .07A .01|1.03 5.41|1.02 4.18|i .53 .56| 54.0 53.1| -.03| SAFEMO5 |

| 26 346344 164995 -.04A .00| .97 -7.99|1.01 2.76|h .52 .57| 59.3 55.6| -.01| SAFEMO1 |

| 34 221590 121865 .63A .00| .97 -7.02|1.01 2.66|g .53 .61| 54.2 52.0| .03| SAFEMO10|

| 47 379812 164703 -.54A .00| .98 -5.73| .91 -9.90|f .60 .54| 62.4 59.2| -.02| SAFBUL9 |

| 42 402941 164652 -.94A .00| .96 -9.90| .88 -9.90|e .56 .50| 66.3 62.5| -.03| SAFBUL1 |

| 37 111186 62666 .42 .01| .94 -9.90| .94 -9.90|d .56 .57| 55.1 50.8| .00| SAFPSF3 |

| 33 67781 42702 .77A .01| .91 -9.90| .92 -9.90|c .53 .59| 55.4 48.9| -.05| SAFEMO9 |

| 31 92407 42979 -.36A .01| .89 -9.90| .87 -9.90|b .56 .53| 60.5 55.4| -.07| SAFEMO7 |

| 45 292260 164544 .63A .00| .87 -9.90| .88 -9.90|a .60 .61| 57.0 51.4| .01| SAFBUL4 |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| MEAN 167412 81113 -.14 .01|1.32 4.6|1.25 4.8| | 54.1 56.4| .00| |

| P.SD 97906.0 45955 .80 .00| .35 8.3| .30 7.8| | 8.7 5.4| .03| |

-------------------------------------------------------------------------------------------------------

**Appendix E4: Technical quality (mean-square error) of 29 Environment items calibrated separately**

-------------------------------------------------------------------------------------------------------

|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |

|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| 71 67141 42955 .84 .01|1.52 9.90|1.66 9.90|A .45 .58| 39.7 48.5| .00| ENVMEN9 |

| 66 134770 59171 -.08A .01|1.20 9.90|1.13 9.90|B .54 .54| 56.3 57.6| -.03| ENVMEN1 |

| 60 233500 122032 .44A .00|1.07 9.90|1.19 9.90|C .31 .58| 54.6 52.7| -.01| ENVINS9 |

| 72 179258 164702 1.90A .00|1.15 9.90|1.18 9.90|D .55 .61| 47.8 47.9| .02| ENVDIS1 |

| 64 85682 59176 1.63 .01|1.11 9.90|1.14 9.90|E .58 .60| 47.8 48.0| .00| ENVINS14|

| 73 139648 59195 -.30A .01|1.12 9.90|1.05 8.06|F .57 .53| 61.3 59.4| -.03| ENVDIS2 |

| 68 125647 62661 -.03A .01|1.11 9.90|1.08 9.90|G .62 .55| 50.2 53.5| .01| ENVMEN4 |

| 58 239475 105536 -.64A .00|1.08 9.90|1.09 9.90|H .42 .51| 57.9 57.4| .00| ENVINS5 |

| 67 129398 62551 -.16A .01|1.05 9.25|1.03 5.03|I .59 .54| 54.5 54.6| .00| ENVMEN3 |

| 75 76011 42915 .43A .01|1.04 5.85|1.04 5.96|J .61 .57| 49.6 51.3| .02| ENVDIS6 |

| 69 92596 42872 -.39A .01|1.02 2.91|1.00 -.57|K .56 .52| 58.5 56.2| .01| ENVMEN6 |

| 74 133614 62583 -.31A .01| .97 -4.59| .95 -7.93|k .56 .53| 59.4 55.7| .00| ENVDIS4 |

| 63 68110 42834 .76A .01| .94 -9.74| .94 -8.78|j .60 .58| 52.5 49.5| .03| ENVINS13|

| 56 304948 121955 -1.03A .01| .92 -9.90| .82 -9.90|i .60 .50| 69.4 63.8| -.02| ENVINS2 |

| 57 155001 59167 -1.15A .01| .92 -9.90| .79 -9.90|h .59 .47| 75.8 69.3| -.04| ENVINS3 |

| 59 158364 62806 -1.36A .01| .90 -9.90| .84 -9.90|g .57 .47| 70.2 64.2| -.02| ENVINS8 |

| 70 127233 59174 .24A .01| .89 -9.90| .87 -9.90|f .61 .56| 60.7 55.7| -.03| ENVMEN7 |

| 76 293549 164928 .63A .00| .88 -9.90| .88 -9.90|e .64 .59| 54.8 51.1| .00| ENVDIS7 |

| 55 326378 164864 .22A .00| .84 -9.90| .87 -9.90|d .49 .57| 59.7 53.8| .00| ENVINS1 |

| 62 67312 42773 .79A .01| .80 -9.90| .82 -9.90|c .63 .58| 56.4 49.4| .03| ENVINS12|

| 61 352276 164618 -.13A .00| .81 -9.90| .80 -9.90|b .66 .55| 62.3 55.6| -.01| ENVINS11|

| 65 90212 42981 -.25A .01| .74 -9.90| .73 -9.90|a .64 .53| 63.5 55.2| .02| ENVINS15|

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| MEAN 162733 82384 .09 .01|1.00 -.3|1.00 -.8| | 57.4 55.0| .00| |

-------------------------------------------------------------------------------------------------------

**Appendix E5: Technical quality (mean-square error) of 13 Bullying items calibrated separately**

-------------------------------------------------------------------------------------------------------

|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |

|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| 48 111398 42839 -1.78 .01|1.82 9.90|1.42 9.90|A .47 .47| 68.5 68.7| .00| SAFBUL10|

| 49 107771 42852 -1.47 .01|1.79 9.90|1.51 9.90|B .48 .50| 61.9 65.3| .00| SAFBUL11|

| 46 137333 62705 -.51 .01|1.70 9.90|1.57 9.90|C .55 .59| 45.4 57.4| .00| SAFBUL5 |

| 43 122153 59169 .56 .01|1.67 9.90|1.56 9.90|D .62 .67| 40.1 54.5| .00| SAFBUL2 |

| 50 129234 59174 .26 .01|1.41 9.90|1.33 9.90|E .58 .65| 51.2 57.0| .00| SAFBUL12|

| 52 112656 62622 .37 .01|1.24 9.90|1.20 9.90|F .67 .64| 43.4 50.9| .00| SAFBUL14|

| 54 91346 42719 -.41 .01|1.09 9.90|1.04 5.48|G .60 .59| 57.1 56.3| .00| SAFBUL16|

| 44 374305 164873 -.45A .00|1.07 9.90|1.04 9.90|f .59 .60| 58.7 58.8| -.03| SAFBUL3 |

| 51 294465 164814 .61A .00| .98 -5.78|1.00 -1.21|e .65 .67| 54.7 51.5| .04| SAFBUL13|

| 53 126323 62642 -.10 .01| .99 -2.04| .96 -7.04|d .60 .62| 57.8 55.1| .00| SAFBUL15|

| 47 379812 164703 -.54A .00| .98 -6.11| .92 -9.90|c .61 .59| 62.8 59.8| -.04| SAFBUL9 |

| 45 292260 164544 .63A .00| .92 -9.90| .96 -9.90|b .63 .67| 56.0 51.5| .04| SAFBUL4 |

| 42 402941 164652 -.94A .00| .95 -9.90| .90 -9.90|a .58 .56| 66.2 63.0| -.06| SAFBUL1 |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| MEAN 206307 96793 -.29 .01|1.28 3.5|1.19 2.8| | 55.7 57.7| .00| |

-------------------------------------------------------------------------------------------------------

**Appendix E6: Technical quality (mean-square error) of 16 reverse-scored items calibrated separately**

-------------------------------------------------------------------------------------------------------

|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |

|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| 49 107771 42852 -1.48 .01|1.70 9.90|1.42 9.90|A .53 .57| 63.9 66.7| .00| SAFBUL11|

| 48 111398 42839 -1.81 .01|1.61 9.90|1.22 9.90|B .56 .55| 70.7 70.6| .00| SAFBUL10|

| 46 137333 62705 -.48 .01|1.52 9.90|1.36 9.90|C .63 .63| 49.3 57.9| .00| SAFBUL5 |

| 35 126168 105625 1.48A .00|1.36 9.90|1.51 9.90|D .53 .67| 44.6 49.8| .12| SAFEMO11|

| 38 156471 62604 -1.38 .01|1.49 9.90|1.28 9.90|E .55 .57| 63.0 66.6| .00| SAFPSF4 |

| 39 107212 42912 -1.42 .01|1.45 9.90|1.21 9.90|F .58 .57| 66.2 66.3| .00| SAFPSF5 |

| 43 122153 59169 .09 .01|1.45 9.90|1.35 9.90|G .70 .74| 48.2 55.7| .00| SAFBUL2 |

| 40 124378 59199 .00 .01|1.44 9.90|1.34 9.90|H .69 .73| 48.7 55.7| .00| SAFPSF7 |

| 50 129234 59174 -.23 .01|1.24 9.90|1.20 9.90|h .67 .72| 52.3 56.7| .00| SAFBUL12|

| 71 67141 42955 .84 .01|1.21 9.90|1.21 9.90|g .63 .66| 46.5 50.6| .00| ENVMEN9 |

| 52 112656 62622 .42 .01|1.15 9.90|1.10 9.90|f .71 .66| 48.6 52.3| .00| SAFBUL14|

| 54 91346 42719 -.39 .01|1.12 9.90|1.09 9.90|e .61 .63| 54.8 56.1| .00| SAFBUL16|

| 41 93477 42820 -.50 .01|1.04 6.24|1.02 2.47|d .61 .62| 58.8 57.5| .00| SAFPSF8 |

| 51 294465 164814 .61A .00| .95 -9.90| .97 -7.29|c .68 .71| 54.8 51.3| -.07| SAFBUL13|

| 37 111186 62666 .47 .01| .87 -9.90| .89 -9.90|b .65 .66| 58.6 52.7| .00| SAFPSF3 |

| 53 126323 62642 -.05 .01| .89 -9.90| .86 -9.90|a .66 .65| 60.4 55.1| .00| SAFBUL15|

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

**Appendix E7: Item category averages: Reverse-scored items (misfit order)**

**|ENTRY DATA SCORE | DATA | ABILITY S.E. INFT OUTF PTMA | |**

**|NUMBER CODE VALUE | COUNT % | MEAN P.SD MEAN MNSQ MNSQ CORR.| ITEM |**

| 40 A 0 0 | 7001 12 | .76 1.01 .01 1.8 2.0 -.25 |SAFPSF7 |

| 1 1 | 9743 16 | 1.12 .89 .01 1.5 1.6 -.15 | |

| 2 2 | 12730 22 | 1.33 .85 .01 1.1 1.1 -.07 | |

| 3 3 | 29725 50 | 1.79 .97 .01 1.3 1.3 .33 | |

| MISSING \*\*\* | 106397 64#| .84 .90 .00 -.30 | |

| | | | |

| 46 B 0 0 | 6253 10 | .26 .88 .01 1.6 1.8 -.22 |SAFBUL5 |

| 1 1 | 9029 14 | .51 .75 .01 1.4 1.4 -.15 | |

| 2 2 | 13965 22 | .73 .75 .01 1.0 1.1 -.07 | |

| 3 3 | 33458 53 | 1.09 .88 .00 1.2 1.2 .29 | |

| MISSING \*\*\* | 102891 62#| 1.19 1.03 .00 .17 | |

| | | | |

| 48 C 0 0 | 1938 5 | .17 .95 .02 2.0 2.2 -.16 |SAFBUL10 |

| 1 1 | 2966 7 | .27 .67 .01 1.4 1.4 -.17 | |

| 2 2 | 5373 13 | .46 .72 .01 1.0 1.0 -.15 | |

| 3 3 | 32562 76 | .99 .93 .01 1.1 1.1 .29 | |

| MISSING \*\*\* | 122757 74#| 1.14 1.00 .00 .13 | |

| | | | |

| 49 D 0 0 | 2091 5 | .28 .94 .02 1.9 2.2 -.14 |SAFBUL11 |

| 1 1 | 3939 9 | .33 .71 .01 1.4 1.4 -.17 | |

| 2 2 | 6634 15 | .53 .73 .01 1.0 1.0 -.14 | |

| 3 3 | 30188 70 | 1.01 .94 .01 1.2 1.2 .29 | |

| MISSING \*\*\* | 122744 74#| 1.14 1.00 .00 .13 | |

| | | | |

| 43 E 0 0 | 7713 13 | .72 .99 .01 1.7 1.9 -.28 |SAFBUL2 |

| 1 1 | 10209 17 | 1.04 .83 .01 1.4 1.4 -.19 | |

| 2 2 | 11797 20 | 1.32 .81 .01 1.0 1.0 -.07 | |

| 3 3 | 29450 50 | 1.85 .96 .01 1.3 1.3 .39 | |

| MISSING \*\*\* | 106427 64#| .84 .90 .00 -.30 | |

| | | | |

| 39 F 0 0 | 2059 5 | .12 .96 .02 1.7 2.0 -.17 |SAFPSF5 |

| 1 1 | 3602 8 | .14 .68 .01 1.2 1.2 -.23 | |

| 2 2 | 8143 19 | .46 .68 .01 .9 .9 -.19 | |

| 3 3 | 29108 68 | 1.08 .91 .01 1.1 1.1 .38 | |

| MISSING \*\*\* | 122684 74#| 1.14 1.00 .00 .13 | |

| | | | |

| 50 G 0 0 | 3910 7 | .84 1.13 .02 2.0 2.4 -.16 |SAFBUL12 |

| 1 1 | 8823 15 | 1.01 .88 .01 1.5 1.5 -.19 | |

| 2 2 | 18912 32 | 1.28 .81 .01 1.0 1.0 -.12 | |

| 3 3 | 27529 47 | 1.81 1.02 .01 1.3 1.3 .32 | |

| MISSING \*\*\* | 106422 64#| .84 .90 .00 -.30 | |

| | | | |

| 38 H 0 0 | 2526 4 | .15 1.06 .02 1.9 2.2 -.16 |SAFPSF4 |

| 1 1 | 5454 9 | .25 .72 .01 1.3 1.3 -.21 | |

| 2 2 | 12855 21 | .54 .71 .01 1.0 .9 -.18 | |

| 3 3 | 41769 67 | 1.06 .85 .00 1.1 1.1 .34 | |

| MISSING \*\*\* | 102992 62#| 1.19 1.03 .00 .17 | |

| | | | |

| 52 J 0 0 | 10196 16 | .21 .82 .01 1.3 1.4 -.32 |SAFBUL14 |

| 1 1 | 13454 21 | .55 .69 .01 1.2 1.2 -.17 | |

| 2 2 | 17714 28 | .88 .71 .01 1.0 1.0 .03 | |

| 3 3 | 21258 34 | 1.30 .90 .01 1.2 1.3 .37 | |

| MISSING \*\*\* | 102974 62#| 1.19 1.03 .00 .17 | |

| | | | |

| 35 K 0 0 | 30020 28 | .44 .81 .00 1.3 1.3 -.28 |SAFEMO11 |

| 1 1 | 35089 33 | .80 .76 .00 1.2 1.2 -.03 | |

| 2 2 | 30469 29 | 1.14 .83 .00 1.2 1.3 .21 | |

| 3 3 | 10047 10 | 1.29 1.21 .01 1.7 2.2 .16 | |

| MISSING \*\*\* | 59971 36#| 1.44 1.02 .00 .29 | |

| 71 N 0 0 | 8319 19 | .26 .79 .01 1.3 1.3 -.30 |ENVMEN9 |

| 1 1 | 11064 26 | .60 .71 .01 1.1 1.2 -.15 | |

| 2 2 | 14639 34 | .97 .77 .01 1.1 1.1 .10 | |

| 3 3 | 8933 21 | 1.45 1.10 .01 1.3 1.4 .34 | |

| MISSING \*\*\* | 122641 74#| 1.14 1.00 .00 .13 | |

| 54 P 0 0 | 2407 6 | .25 1.04 .02 1.6 1.9 -.15 |SAFBUL16 |

| 1 1 | 6470 15 | .32 .69 .01 1.1 1.1 -.24 | |

| 2 2 | 16650 39 | .70 .71 .01 .9 .9 -.12 | |

| 3 3 | 17192 40 | 1.25 .99 .01 1.2 1.2 .36 | |

| MISSING \*\*\* | 122877 74#| 1.14 1.00 .00 .13 | |

| | | | |

**Appendix E7: Item category averages: Reverse-scored items (misfit order) continued**

**|ENTRY DATA SCORE | DATA | ABILITY S.E. INFT OUTF PTMA | |**

**|NUMBER CODE VALUE | COUNT % | MEAN P.SD MEAN MNSQ MNSQ CORR.| ITEM |**

| 41 Q 0 0 | 2005 5 | .28 1.00 .02 1.6 1.9 -.13 |SAFPSF8 |

| 1 1 | 5465 13 | .26\* .70 .01 1.1 1.1 -.24 | |

| 2 2 | 18038 42 | .68 .72 .01 .9 .9 -.15 | |

| 3 3 | 17312 40 | 1.26 .99 .01 1.1 1.1 .37 | |

| MISSING \*\*\* | 122776 74#| 1.14 1.00 .00 .13 | |

| 51 S 0 0 | 16145 10 | .36 1.00 .01 1.4 1.6 -.23 |SAFBUL13 |

| 1 1 | 39385 24 | .70 .81 .00 1.2 1.3 -.20 | |

| 2 2 | 72772 44 | 1.09 .81 .00 1.1 1.1 .03 | |

| 3 3 | 36512 22 | 1.70 1.08 .01 1.2 1.3 .35 | |

| MISSING \*\*\* | 782 0#| .31 1.23 .04 -.05 | |

| 53 Y 0 0 | 3793 6 | .15 1.00 .02 1.4 1.6 -.20 |SAFBUL15 |

| 1 1 | 10143 16 | .40 .71 .01 1.1 1.1 -.22 | |

| 2 2 | 29938 48 | .79 .71 .00 1.0 .9 -.05 | |

| 3 3 | 18768 30 | 1.30 .95 .01 1.2 1.2 .34 | |

| MISSING \*\*\* | 102954 62#| 1.19 1.03 .00 .17 | |

| 37 0 0 | 5564 9 | .13 .93 .01 1.3 1.3 -.25 |SAFPSF3 |

| 1 1 | 14362 23 | .46 .71 .01 1.0 1.0 -.24 | |

| 2 2 | 31396 50 | .93 .72 .00 1.0 1.0 .09 | |

| 3 3 | 11344 18 | 1.45 .98 .01 1.2 1.2 .32 | |

| MISSING \*\*\* | 102930 62#| 1.19 1.03 .00 .17 | |

**Appendix F:** **Winsteps residual analyses output**

--------------------------------------------------------------------------------------

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

Eigenvalue Observed Expected

Total raw variance in observations = 120.8266 100.0% 100.0%

Raw variance explained by measures = 44.8266 37.1% 41.0%

Raw variance explained by persons = 25.9884 21.5% 23.8%

Raw Variance explained by items = 18.8382 15.6% 17.2%

Raw unexplained variance (total) = 76.0000 62.9% 100.0% 59.0%

Unexplned variance in 1st contrast = 3.4010 2.8% 4.5%

Unexplned variance in 2nd contrast = 2.7261 2.3% 3.6%

Unexplned variance in 3rd contrast = 2.4312 2.0% 3.2%

Unexplned variance in 4th contrast = 2.3388 1.9% 3.1%

Unexplned variance in 5th contrast = 1.8956 1.6% 2.5%

**STANDARDIZED RESIDUAL CONTRAST 1 PLOT (All labeled items are reverse scored items)**

-5 -4 -3 -2 -1 0 1 2 3 4 5 6

-+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+- COUNT CLUSTER

.6 + | **BUL13** + 1 1

C | | |

O .5 + **BUL15** + 1 1

N | | **BUL14 PSF3** | 2 1

T .4 + **BUL5** | + 1 1

R | PSF4 | | 1 1

A .3 + BUL10 BUL11 PSF8 BUL16 | J + 5 1

S | PSF5 | | 1 1

T .2 + M|BUL12 PSF7 EMO11 + 4 1

| | QTRS | 4 2

1 .1 + U | + 1 2

| W V X |Z Y | 5 2

L .0 +------------------------1--13111--1--------------------------------+ 9 2

O | 1111 1 | 11 | 7 2

A -.1 + 1 121| 1z1 1 + 9 2

D | qyv wx r|s upt | 10 3

I -.2 + m oij lnk + 7 3

N | | h | 1 3

G -.3 + d gc e f + 5 3

| b| a | 2 3

-.4 + | +

-+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-

-5 -4 -3 -2 -1 0 1 2 3 4 5 6

ITEM MEASURE

COUNT: 1 535226875425972 111

**Approximate relationships between the PERSON measures1**

PCA ITEM Pearson Disattenuated Pearson+Extr Disattenuated+Extr

Contrast Clusters Correlation Correlation Correlation Correlation

1 1 - 3 0.3512 0.4359 0.3558 0.4415

1 1 - 2 0.5341 0.6704 0.5374 0.6745

1 2 - 3 0.7571 0.8987 0.7595 0.9016

----------------------------------------------------- ----------------------------------------------

|CON- | | INFIT OUTFIT| ENTRY | | | INFIT OUTFIT| ENTRY |

| TRAST|LOADING|MEASURE MNSQ MNSQ |NUMBER ITEM | |LOADING|MEASURE MNSQ MNSQ |NUMBER ITEM |

|------+-------+-------------------+----------------| |-------+-------------------+----------------|

| 1 | .59 | .61 1.33 1.39 |A 51 **SAFBUL13** | | -.35 | .99 .99 1.00 |a 11 ENGPAR4 |

| 1 | .48 | -.05 1.28 1.29 |B 53 **SAFBUL15** | | -.35 | -.13 .90 .88 |b 61 ENVINS11 |

| 1 | .46 | .39 1.80 1.80 |C 52 **SAFBUL14** | | -.32 | .07 .87 .86 |c 29 SAFEMO5 |

| 1 | .43 | .44 1.12 1.15 |D 37 **SAFPSF3** | | -.31 | -1.03 1.02 .91 |d 56 ENVINS2 |

| 1 | .41 | -.44 2.20 2.19 |E 46 **SAFBUL**5 | | -.31 | .63 .98 .98 |e 76 ENVDIS7 |

| 1 | .33 | -1.28 1.85 1.81 |F 38 **SAFPSF4**  | | -.29 | .97 .93 .95 |f 17 ENGPAR10 |

| 1 | .30 | -1.65 2.19 2.10 |G 48 **SAFBUL10** | | -.28 | -.17 .78 .77 |g 24 ENGREL13 |

| 1 | .30 | -.33 1.48 1.53 |H 54 **SAFBUL16** | | -.23 | 1.90 1.33 1.39 |h 72 ENVDIS1 |

| 1 | .29 | -.44 1.38 1.41 |I 41 **SAFPSF8** | | -.22 | -.04 .86 .89 |i 26 SAFEMO1 |

| 1 | .28 | .75 .66 .67 |J 20 ENGREL1 | | -.21 | -.03 1.22 1.18 |j 68 ENVMEN4 |

| 1 | .28 | -1.35 2.17 2.16 |K 49 **SAFBUL11** | | -.20 | .79 .95 .97 |k 62 ENVINS12 |

| 1 | .26 | -1.30 1.90 1.83 |L 39 **SAFPSF5** | | -.20 | .76 1.10 1.11 |l 63 ENVINS13 |

| 1 | .22 | -.22 .86 .85 |M 5 ENGCLC5 | | -.19 | -.47 .80 .80 |m 23 ENGREL6 |

| 1 | .21 | 1.48 1.49 1.68 |N 35 **SAFEMO11** | | -.19 | .84 1.39 1.40 |n 28 SAFEMO4 |

| 1 | .21 | .16 1.84 1.89 |O 50 **SAFBUL12** | | -.18 | -.16 1.18 1.16 |o 67 ENVMEN3 |

| 1 | .20 | .35 2.23 2.24 |P 40 **SAFPSF7** | | -.17 | .61 1.20 1.17 |p 14 ENGPAR7 |

----------------------------------------------------- ----------------------------------------------

1Bolded items form the 1st contrast and are all reverse-scored items.

**Appendix G: Measure order of 76-item VOCAL scale**

-------------------------------------------------------------------------------------------------------

|ENTRY TOTAL TOTAL MODEL| INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH| | |

|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%|DISPLACE| ITEM |

|------------------------------------+----------+----------+-----------+-----------+--------+---------|

| 72 179258 164702 1.90A .00|1.33 9.90|1.39 9.90| .44 .58| 44.3 47.2| -.01| ENVDIS1 |

| 64 85682 59176 1.64 .01|1.30 9.90|1.35 9.90| .48 .58| 44.1 48.4| .00| ENVINS14|

| 35 126168 105625 1.48A .00|1.49 9.90|1.68 9.90| .33 .55| 39.2 45.8| -.01| SAFEMO11|

| 11 261745 164675 .99A .00| .99 -3.21|1.00 -1.17| .54 .57| 50.7 48.8| -.01| ENGPAR4 |

| 17 93322 62734 .97A .01| .93 -9.90| .95 -9.43| .47 .54| 50.8 47.6| -.01| ENGPAR10|

| 22 110102 59166 .86A .01| .66 -9.90| .67 -9.90| .57 .57| 65.5 52.0| .00| ENGREL4 |

| 28 97433 62610 .84A .01|1.39 9.90|1.40 9.90| .53 .54| 39.4 48.2| .00| SAFEMO4 |

| 71 67141 42955 .81 .01|1.53 9.90|1.58 9.90| .43 .54| 38.7 48.2| .00| ENVMEN9 |

| 62 67312 42773 .79A .01| .95 -8.45| .97 -5.54| .53 .54| 50.8 48.5| .00| ENVINS12|

| 33 67781 42702 .77A .01| .81 -9.90| .82 -9.90| .57 .54| 55.3 48.5| .00| SAFEMO9 |

| 63 68110 42834 .76A .01|1.10 9.90|1.11 9.90| .49 .54| 46.6 48.5| .00| ENVINS13|

| 20 283306 164952 .75A .00| .66 -9.90| .67 -9.90| .56 .56| 62.7 50.1| -.01| ENGREL1 |

| 32 114929 59168 .70A .01|1.29 9.90|1.26 9.90| .56 .56| 47.2 52.7| .00| SAFEMO8 |

| 19 70987 42821 .64A .01|1.11 9.90|1.14 9.90| .43 .54| 47.1 49.2| .00| ENGPAR12|

| 34 221590 121865 .63A .00| .85 -9.90| .85 -9.90| .56 .56| 57.2 51.3| .00| SAFEMO10|

| 45 292260 164544 .63A .00| .85 -9.90| .86 -9.90| .60 .56| 55.8 50.8| .00| SAFBUL4 |

| 76 293549 164928 .63A .00| .98 -5.06| .98 -7.36| .59 .56| 51.6 50.8| .00| ENVDIS7 |

| 3 103805 62113 .63 .01|1.70 9.90|1.81 9.90| .25 .53| 39.5 49.5| .00| ENGCLC3 |

| 15 116919 59177 .62A .01|1.16 9.90|1.14 9.90| .46 .56| 51.9 53.2| .01| ENGPAR8 |

| 14 117259 59168 .61A .01|1.20 9.90|1.17 9.90| .55 .56| 48.7 53.2| .00| ENGPAR7 |

| 51 294465 164814 .61A .00|1.33 9.90|1.39 9.90| .41 .56| 46.9 51.0| .00| SAFBUL13|

| 12 108596 62797 .52A .01|1.09 9.90|1.11 9.90| .39 .53| 49.6 50.2| .00| ENGPAR5 |

| 60 233500 122032 .44A .00|1.13 9.90|1.26 9.90| .21 .55| 54.2 52.4| .00| ENVINS9 |

| 37 111186 62666 .44 .01|1.12 9.90|1.15 9.90| .44 .52| 51.6 50.9| .00| SAFPSF3 |

| 43 122153 59169 .43 .01|2.16 9.90|2.15 9.90| .42 .55| 32.6 54.1| .00| SAFBUL2 |

| 75 76011 42915 .43A .01|1.13 9.90|1.14 9.90| .56 .53| 46.2 50.9| .00| ENVDIS6 |

| 52 112656 62622 .39 .01|1.80 9.90|1.80 9.90| .44 .52| 33.7 51.2| .00| SAFBUL14|

| 40 124378 59199 .35 .01|2.23 9.90|2.24 9.90| .36 .54| 33.7 54.5| .00| SAFPSF7 |

| 70 127233 59174 .24A .01| .88 -9.90| .85 -9.90| .63 .54| 60.0 55.0| .00| ENVMEN7 |

| 55 326378 164864 .22A .00| .84 -9.90| .86 -9.90| .45 .54| 59.9 53.3| .00| ENVINS1 |

| 50 129234 59174 .16 .01|1.84 9.90|1.89 9.90| .34 .53| 44.5 55.5| .00| SAFBUL12|

| 18 82115 42980 .15A .01|1.23 9.90|1.23 9.90| .46 .51| 45.0 52.7| .00| ENGPAR11|

| 29 122626 62609 .07A .01| .87 -9.90| .86 -9.90| .63 .51| 56.4 53.0| .00| SAFEMO5 |

| 68 125647 62661 -.03A .01|1.22 9.90|1.18 9.90| .57 .51| 48.1 53.4| .00| ENVMEN4 |

| 26 346344 164995 -.04A .00| .86 -9.90| .89 -9.90| .57 .53| 61.6 54.8| .00| SAFEMO1 |

| 53 126323 62642 -.05 .01|1.28 9.90|1.29 9.90| .40 .50| 52.0 53.7| .00| SAFBUL15|

| 66 134770 59171 -.08A .01|1.27 9.90|1.20 9.90| .52 .52| 55.2 57.2| .01| ENVMEN1 |

| 27 135280 59186 -.10A .01|1.09 9.90|1.01 2.14| .63 .52| 59.7 57.3| .01| SAFEMO3 |

| 61 352276 164618 -.13A .00| .90 -9.90| .88 -9.90| .62 .52| 59.3 55.3| .00| ENVINS11|

| 4 87939 42837 -.14 .01|1.64 9.90|1.67 9.90| .40 .50| 40.9 54.4| .00| ENGCLC4 |

| 67 129398 62551 -.16A .01|1.18 9.90|1.16 9.90| .52 .50| 51.0 54.1| .00| ENVMEN3 |

| 24 218132 105459 -.17A .00| .78 -9.90| .77 -9.90| .62 .50| 61.8 54.2| .00| ENGREL13|

| 5 220622 105472 -.22A .00| .86 -9.90| .85 -9.90| .55 .50| 60.5 54.5| .00| ENGCLC5 |

| 65 90212 42981 -.25A .01| .82 -9.90| .80 -9.90| .59 .49| 59.9 54.8| .00| ENVINS15|

| 73 139648 59195 -.30A .01|1.18 9.90|1.13 9.90| .55 .51| 59.6 58.8| .01| ENVDIS2 |

| 74 133614 62583 -.31A .01|1.06 9.48|1.03 5.77| .51 .49| 56.3 54.8| .00| ENVDIS4 |

| 8 140245 59186 -.32 .01|1.67 9.90|1.64 9.90| .38 .50| 51.6 59.2| .00| ENGPAR1 |

| 54 91346 42719 -.33 .01|1.48 9.90|1.53 9.90| .38 .49| 47.7 55.1| .00| SAFBUL16|

| 31 92407 42979 -.36A .01| .89 -9.90| .87 -9.90| .57 .49| 59.9 55.4| .00| SAFEMO7 |

| 16 141117 59201 -.37A .01|1.36 9.90|1.38 9.90| .31 .50| 51.8 59.4| .01| ENGPAR9 |

| 6 141284 59158 -.38A .01|1.30 9.90|1.29 9.90| .42 .50| 57.2 59.9| .01| ENGCLC6 |

| 69 92596 42872 -.39A .01|1.06 9.09|1.04 5.54| .53 .48| 55.9 55.5| .01| ENVMEN6 |

| 41 93477 42820 -.44 .01|1.38 9.90|1.41 9.90| .38 .48| 51.9 55.8| .00| SAFPSF8 |

| 46 137333 62705 -.44 .01|2.20 9.90|2.19 9.90| .33 .48| 34.7 55.4| .00| SAFBUL5 |

| 44 374305 164873 -.45A .00|1.11 9.90|1.11 9.90| .56 .50| 58.3 57.3| .00| SAFBUL3 |

| 23 232709 105781 -.47A .00| .80 -9.90| .80 -9.90| .54 .48| 62.2 55.7| .00| ENGREL6 |

| 13 144689 59192 -.54A .01|1.06 9.36| .99 -.85| .52 .49| 61.6 61.5| .00| ENGPAR6 |

| 47 379812 164703 -.54A .00| .97 -9.15| .90 -9.90| .64 .50| 62.7 58.0| .00| SAFBUL9 |

| 9 140206 62276 -.59A .01|1.24 9.90|1.18 9.90| .51 .48| 56.1 56.0| .01| ENGPAR2 |

| 58 239475 105536 -.64A .00|1.12 9.90|1.14 9.90| .36 .47| 55.7 56.5| .00| ENVINS5 |

| 36 150540 59204 -.86A .01|1.20 9.90|1.09 9.90| .51 .47| 67.2 65.5| .00| SAFPSF1 |

| 1 401337 164624 -.91A .00|1.09 9.90|1.01 1.73| .54 .47| 62.6 61.1| .00| ENGCLC1 |

| 25 251296 105544 -.94A .01| .90 -9.90| .85 -9.90| .59 .45| 66.2 58.5| .01| ENGREL14|

| 42 402941 164652 -.94A .00| .96 -9.60| .89 -9.90| .59 .47| 67.2 61.4| .00| SAFBUL1 |

| 10 103150 42980 -1.00 .01|1.27 9.90|1.21 9.90| .47 .44| 57.3 59.0| .00| ENGPAR3 |

| 56 304948 121955 -1.03A .01|1.02 4.68| .91 -9.90| .56 .46| 67.4 63.4| .00| ENVINS2 |

| 7 255949 105486 -1.07A .01|1.03 6.80| .99 -2.47| .45 .44| 60.6 59.7| .00| ENGCLC7 |

| 57 155001 59167 -1.15A .01| .98 -3.57| .82 -9.90| .58 .44| 75.3 69.5| .01| ENVINS3 |

| 30 105128 42832 -1.16 .01|1.57 9.90|1.48 9.90| .39 .43| 55.7 60.5| .00| SAFEMO6 |

| 2 155583 59165 -1.19A .01|1.13 9.90| .90 -9.90| .59 .44| 75.7 69.9| .01| ENGCLC2 |

| 21 156891 59200 -1.27A .01|1.12 9.90| .91 -9.90| .55 .43| 75.2 71.2| .01| ENGREL3 |

| 38 156471 62604 -1.28 .01|1.85 9.90|1.81 9.90| .35 .43| 56.6 61.7| .00| SAFPSF4 |

| 39 107212 42912 -1.30 .01|1.90 9.90|1.83 9.90| .37 .42| 57.3 61.8| .00| SAFPSF5 |

| 49 107771 42852 -1.35 .01|2.17 9.90|2.16 9.90| .28 .42| 54.9 62.4| .00| SAFBUL11|

| 59 158364 62806 -1.36A .01| .98 -3.61| .90 -9.90| .53 .43| 67.3 62.9| .01| ENVINS8 |

| 48 111398 42839 -1.65 .01|2.19 9.90|2.10 9.90| .29 .39| 64.8 66.6| .00| SAFBUL10|

**Appendix H: Item prompts by dimension**

**Appendix H1: Engagement items (Stem: Think of the last 30 days)**

| **Grade** | **Item code1** | **Cultural and Linguistic Competence domain** i**tem prompts** |
| --- | --- | --- |
| 8, 10 | ENGCLC12 | Adults working at this school treat all students respectfully, regardless of a student's race, culture, family income, religion, sex, or sexual preference. |
| 5 | ENGCLC12 | Adults working at this school treat all students respectfully. |
| 5 | ENGCLC2 | Teachers at this school accept me for who I am. |
| 8 | ENGCLC32 | My textbooks or class materials include people and examples that reflect my race, cultural background and/or identity. |
| 10 | ENGCLC4 | I am encouraged to take upper level courses (honors, AP). |
| 8, 10 | ENGCLC52 | Students from different backgrounds respect each other in our school, regardless of their race, culture, family income, religion, sex, or sexual preference. |
| 5 | ENGCLC6 | Students like to have friends who are different (for example, boys and girls, rich and poor, or classmates of different color). |
| 8, 10 | ENGCLC7 | Students are open to having friends who come from different backgrounds (for example, friends from different races, cultures, family incomes, or religions, or friends of a different sex or sexual preference). |
| **Grade** | **Item code1** | **Relationships domain** i**tem prompts** |
| 5, 8, 10 | ENGREL12 | Students respect one another. |
| 5 | ENGREL32 | My teachers care about me as a person. |
| 5 | ENGREL42 | Students at my school get along well with each other. |
| 8, 10 | ENGREL62 | Teachers are available when I need to talk with them. |
| 8, 10 | ENGREL13 | Adults at our school are respectful to student ideas even if the ideas expressed are different from their own. |
| 8, 10 | ENGREL14 | My teachers promote respect among students. |

1Items in bold are reverse-scored items;2Item taken from or adapted from [EDSCLS survey](https://nces.ed.gov/surveys/edscls/questionnaires.asp);

**Appendix H1: Engagement items continued**

| **Grade** | **Item code1** | **Participation domain** i**tem prompts** |
| --- | --- | --- |
| 5 | ENGPAR12 | I get the chance to take part in school events (for example, science fairs, art or music shows). |
| 8 | ENGPAR2 | My parents feel respected when they participate at our school (e.g., at parent-teacher conferences, open houses). |
| 10 | ENGPAR32 | I feel welcome to participate in extra-curricular activities offered through my school, such as, school clubs or organizations, musical groups, sports teams, student council, or any other extra-curricular activities. |
| 5, 8, 10 | ENGPAR4 | My teachers use my ideas to help my classmates learn. |
| 8 | ENGPAR5 | I have a choice in how I show my learning (e.g., write a paper, prepare a presentation, make a video). |
| 5 | ENGPAR6 | My teachers will explain things in different ways until I understand. |
| 5 | ENGPAR7 | When I need help, my teachers use my interests to help me learn. |
| 5 | ENGPAR8 | My teachers ask me to share what I have learned in a lesson. |
| 5 | ENGPAR9 | When I am stuck, my teachers want me to try again before they help me. |
| 8 | ENGPAR10 | In my class, my teachers use students' interests to plan class activities. |
| 10 | ENGPAR11 | In at least two of my academic classes, I am allowed to work on assignments that interest me personally. |
| 10 | ENGPAR12 | If I finish my work early, I have an opportunity to do more challenging work. |

**1**Items in bold are reverse-scored items;2Item taken from or adapted from [EDSCLS survey](https://nces.ed.gov/surveys/edscls/questionnaires.asp);

**Appendix H2: Safety items (Stem: Think of the last 30 days)**

| **Grade** | **Item code1** | **Emotional safety domain item prompts** |
| --- | --- | --- |
| 5, 8, 10 | SAFEMO1 | Teachers support (help) students who come to class upset. |
| 5 | SAFEMO32 | I am happy to be at our school. |
| 8 | SAFEMO4 | I feel comfortable reaching out to teachers/counselors for emotional support if I need it. |
| 8 | SAFEMO5 | Teachers and adults are interested in my well-being beyond just my class work. |
| 10 | SAFEMO6 | I have a group of friends I can rely on to help me when I feel down (sad). |
| 10 | SAFEMO7 | I feel as though I belong to my school community. |
| 5 | SAFEMO82 | I feel comfortable talking to my teacher(s) about something that is bothering me. |
| 10 | SAFEMO9 | Students at school try to work out their problems with other students in a respectful way. |
| 5, 8 | SAFEMO10 | Students will help other students if they are upset, even if they are not close friends. |
| 8, 10 | **SAFEMO11** | Because I worry about my grades, it is hard for me to enjoy school. |
| **Grade** | **Item code1** | **Physical safety domain item prompts** |
| 5 | SAFPSF12 | I feel safe at our school. |
| 8 | **SAFPSF3**1,2 | Students at this school damage and/or steal other students' property. |
| 8 | **SAFPSF4**1,2 | I have seen students with weapons at our school. |
| 10 | **SAFPSF5**1,2 | I have stayed at home (or avoided school) because I did not feel safe at my school. |
| 5 | **SAFPSF7** | In the last month, I have seen more than one physical fight at my school. |
| 10 | **SAFPSF8** | Students are sexually harassed at my school (for example, bothered by unwanted touching and/or indecent name-calling). |

**1**Items in bold are reverse-scored items;2Item taken from or adapted from [EDSCLS survey](https://nces.ed.gov/surveys/edscls/questionnaires.asp);

**Appendix H21: Safety items continued**

| **Grade** | **Item code1** | **Bullying/Cyber-bullying domain item prompts** |
| --- | --- | --- |
| 5, 8, 10 | SAFBUL1 | If I tell a teacher or other adult that someone is being bullied, the teacher/adult will do something to help. |
| 5 | SAFBUL2 | I have been punched or shoved by other students **more than once** in the school or in the playground. |
| 5, 8, 10 | SAFBUL3 | Teachers **don't** let students pick on other students in class or in the hallways. |
| 5, 8, 10 | SAFBUL42 | Students at this school try to stop bullying when they see it happening. |
| 8 | **SAFBUL5**1,2 | Students have spread rumors or lies about me **more than once** on social media. |
| 5, 8, 10 | SAFBUL9 | Teachers, students, andthe principal work together in our school to prevent (stop) bullying. |
| 10 | **SAFBUL10**1,2 | I have been teased or picked **on more than once** because of my real or perceived sexual preference. |
| 10 | **SAFBUL11**1,2 | I have been teased or picked on **more than once** because of my race or ethnicity. |
| 5 | **SAFBUL12** | In my school, older students scare or pick on younger students. |
| 5, 8, 10 | **SAFBUL13** | In my school, groups of students tease or pick on one student. |
| 8 | **SAFBUL14** | I have been called names or made fun of by other students **more than once** in school. |
| 8 | **SAFBUL15** | In my school, bigger students taunt or pick on smaller students. |
| 10 | **SAFBUL16**1,2 | Students with learning or physical difficulties are teased or picked on at my school. |

**1**Items in bold are reverse-scored items;2Item taken from or adapted from [EDSCLS survey](https://nces.ed.gov/surveys/edscls/questionnaires.asp);

**Appendix H31: Environment items (Stem: Think of the last 30 days)**

| **Grade** | **Item code1** | **Instructional environment domain item prompts** |
| --- | --- | --- |
| 5, 8, 10 | ENVINS1 | Students help each other learn without having to be asked by the teacher. |
| 5, 8 | ENVINS2 | My teachers are proud of me when I work hard in school. |
| 5 | ENVINS3 | My teachers help me succeed with my school work when I need help. |
| 8, 10 | ENVINS5 | My teachers set high expectations for my work. |
| 8 | ENVINS8 | My teachers believe that **all** students can do well in their learning. |
| 5, 8 | ENVINS9 | My schoolwork is challenging (hard) but not too difficult. |
| 5, 8, 10 | ENVINS11 | My teachers support me even when my work is not my best. |
| 10 | ENVINS12 | The things I am learning in school are relevant (important) to me. |
| 10 | ENVINS133 | Teachers ask students for feedback on their classroom instruction. |
| 5 | ENVINS143 | When I am home, I like to learn more about what I did in school. |
| 10 | ENVINS15 | My teachers inspire confidence in my ability to be ready for college or career. |
| **Grade** | **Item code1** | **Mental health environment domain item prompts** |
| 5 | ENVMEN1 | In school, I learn how to control my feelings when I am angry or upset. |
| 8 | ENVMEN3 | Our school offers guidance to students on how to mediate (settle) conflicts by themselves. |
| 8 | ENVMEN4 | If I need help with my emotions (feelings), **effective** help is available at my school. |
| 10 | ENVMEN6 | I have access to **effective** help at school if I am struggling emotionally or mentally. |
| 5 | ENVMEN72 | At our school, students learn to care about other students' feelings. |
| 10 | **ENVMEN91** | The level of pressure I feel at school to perform well is unhealthy. |
| **Grade** | **Item code1** | **Discipline environment domain item prompts** |
| 5, 8, 10 | ENVDIS1 | Students have a voice in deciding school rules. |
| 5 | ENVDIS2 | School rules are fair for all students. |
| 8 | ENVDIS42 | School staff are consistent when enforcing rules in school. |
| 10 | ENVDIS6 | The consequences for the same inappropriate behavior (e.g., disrupting the class) are the same, no matter who the student is. |
| 5, 8, 10 | ENVDIS7 | Teachers give students a chance to explain their behavior when they do something wrong. |

**1**Items in bold are reverse-scored items;2Item taken from or adapted from [EDSCLS survey](https://nces.ed.gov/surveys/edscls/questionnaires.asp); 3Item taken from or adapted from [Panorama Education student survey](https://www.panoramaed.com/panorama-student-survey)

**Appendix I:**

**Person Reliability of VOCAL scale, grade-level VOCAL scales and dimension sub-scales**

**Overall School Climate (persons = 165,587; items = 76)1**

|  | Person Separation Reliability (PSR)2 | Person Separation Index (PSI: G) | Person Strata (H) | Mean ±SD3 |
| --- | --- | --- | --- | --- |
| Real – Model | 0.91 – 0.93 | 3.11 – 3.54 | 4.5 – 5.0 | 1.06 ± 0.99 |

**Grade 5 items (persons = 59,216; items = 36)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.90 – 0.92 | 2.92 – 3.36 | 4.2 – 4.8 | 1.46 ± 1.01 |

**Grade 8 items (persons =62,857; items = 38)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.90 – 0.92 | 3.02 – 3.40 | 4.4 – 4.9 | 0.84 ± 0.88 |

**Grade 10 items (43,514; items = 38)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.89 – 0.92 | 2.91 – 3.32 | 4.2 – 4.8 | 0.83 ± 0.93 |

**Engagement (persons = 165,482; items = 25)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.77 – 0.80 | 1.81 – 1.99 | 2.7 – 3.0 | 1.09 ± 1.12 |

**Safety items (persons = 165,481; items = 29)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.81 – 0.85 | 2.05 – 2.35 | 3.1 – 3.5 | 1.12 ± 1.22 |

**Environment items (persons = 165,469; items = 22)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.76 – 0.80 | 1.78 – 2.01 | 2.7 – 3.0 | 1.08 ± 1.11 |

**Bullying/Cyberbullying items (persons = 165,349; items = 13)1**

|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD |
| --- | --- | --- | --- | --- |
| Real - Model | 0.71 – 0.76 | 1.58 – 1.76 | 2.4 – 2.7 | 1.20 ± 1.46 |

**1**13 common items: grade 5, 8, and 10; 7 common items: grade 8 and 10; 4 common items: grade 5 and 8

**2**Real person separation reliability: lower bound of reliability; Model PSR: upper bound; **3**SD: Standard Deviation.

**Appendix J1: DIF Plot: Economically disadvantaged (ECODIS)**

**Appendix J2: DIF Plot: Students with disabilities (SWD)**

**Appendix J3: DIF Plot: English Learner (EL)**

**Appendix K: Transformation of logit scores**

To transform student-level person measures into interpretable school-level scores, the following steps were taken:

1. The school climate person measures were exported out from Winsteps based on the joint calibration of all students (all students from across the three grades included),
2. Each person’s logit measure was standardized by subtracting the mean of the overall school climate measure from each students’ score and dividing by the standard deviation of the overall school climate measure

sclstd=

where sclstd is the person’s standardized school climate measure

1. The standardized estimates were then multiplied by 20 and 50 was added to each individual score.

As a result of this process, student scores were centered at 50 with a standard deviation of 20. Before aggregation to the school-level, student scores were truncated to range from 1 to 99. As a result, school-level scores had a mean of 50.05 and a standard deviation of 12.83. A similar process was used for each dimension score.