|  |  |  |
| --- | --- | --- |
| ESE logo | | |
|  | 2022 Views of Climate and Learning (VOCAL) Deeper Learning Items: Validity Addendum | |
|  | |
| February, 2023 | |
| 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.  © 2023 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](http://www.doe.mass.edu)  State Seal of Massachusetts | | | |

**Table of Contents**

1. **Purpose of this report**…………………………………………………………….…... 1

**2.** **School climate construct** **and raw response scoring**……………………………...… 2

**3. 2022 survey improvements**……………………………………………………………3

**4. Profile of 2022 Respondents**……………………………………………………..…... 5

**5. Data analyses procedures**……………………………………………………………. 6

**6. Validity evidence for 2012 VOCAL survey** …………………………………………7

6.1. Validity framework...…………………………………………….......................... 7

6.2. Content validity.. …………………………………………………………………9

6.3. Structural validity………………………………………………………………… 10

6.4. Substantive validity……………………………………………………………….13

6.5. Generalizability…………………………………………………………………... 14

6.6. External validity…………………………………………………………………..19

6.7. Consequential validity…………………………………………………………… 23

**6. VOCAL report conclusion**…………………………………………………………… 25

**References**…………………………………………………………………………………. 28

**Validity appendices**……..……………………………………………………..…………..30

A. 2022 VOCAL survey specifcation by grade……………………………………….. 30

B. Guide for evaluating Rasch model validity data……..…………………………..…31

C1. **Content**: Technical quality of 113-item VOCAL scale ……….………………. 32

C2**. Content**: Technical quality of 42-item Engagement dimension scale ……….….. 33

C3. **Content**: Technical quality of 34-item Safety dimension scale ….……………… 34

C4. **Content**: Technical Quality of 37-item Environment dimension scale…..……… 35

C5. **Content:** Technical Quality of 24-item Participation topic scale ………..……… 36

C6. **Content**: Technical Quality of 17-item Bullying topic scale …..……………...… 36

C7. **Content**: Technical Quality of 21-item Instructional topic scale …………...…… 37

D1. **Structural:** Residual analyses of 113-item VOCAL scale ……………………… 38

D2**. Structural:** Residual analyses of dimension scales……………………………… 38

D2. **Structural:** Residual analyses of topic scales……………….……........................ 38

E. **Structural:** Pearson correlations between dimension/topic scores……………… 39

F. **Substantive:** Rate scale structure of 113-item VOCAL scale ……………….…… 40

G1. **Substantive:** 113-item-variable map with dimensions color coded…………….... 41

G2. **Substantive:** 113-item-variable map with topics color coded ……….…………… 42

G3. **Substantive:** Item-threshold map for 113-item VOCAL scale…............................ 43

H1. **Generalizability:** Person separation reliability of 113-item jointly calibrated VOCAL scale and of individual grade-level scales………….…………………… 44

H2. **Generalizability:** Person separation reliability of three-dimension scales............. 45

H3. **Generalizability:** Person separation reliability of three-topic scales……….......... 45

I1. **Generalizability:** Differential Item Functioning by gender status……….............. 46

I2. **Generalizability:** Differential Item Functioning by race/ethnicity status….…..... 47

I3. **Generalizability:** Differential Item Functioning by low-income status…….….... 49

I4. **Generalizability:** Differential Item Functioning by English learner status…........ 50

I5. **Generalizability:** Differential Item Funct. by students with disabilities status….. 51

J1. **External validity:** School-level VOCAL correlations: Behavioral indicators……. 52

J2. **External validity:** School-level VOCAL correlations: Academic achievement and growth scores and overall VOCAL scores (by grade)………………….….........… 52

J3. **External validity:** School-level VOCAL correlations: English Language Arts

scaled scores and VOCAL dimension scores (by grade)………………...……...… 53

J4. **External validity:** School-level VOCAL correlations: English Language Arts

growth scores and VOCAL dimension scores (by grade)…………………….…… 53

J5. **External validity:** School-level VOCAL correlations: Mathematics scaled scores

and VOCAL dimension scores (by grade)…….……………………………...….… 53

J6. **External validity:** School-level VOCAL correlations: Mathematics growth

scores and VOCAL dimension scores (by grade)………………………………...... 54

**1. Purpose of this report**

This report serves as an addendum to the reliability and validity evidence provided in the [2018 Views of Climate and Learning (VOCAL) validity study](http://www.doe.mass.edu/research/vocal/2018/validity-study.docx) (DESE, 2019a) and in the [2019 Views of Climate and Learning (VOCAL) grade 4 validity study addendum](https://www.doe.mass.edu/research/vocal/default.html) (DESE2023a). The VOCAL survey was developed by the Massachusetts Department of Elementary and Secondary Education (DESE, 2019a); the survey is predicated on the United States Department of Education’s conceptual framework for school climate (USED, 2019) and is administered in grades 4, 5, 8, and 10. This report complements the validity work reported previously ([[DESE 2019](https://www.doe.mass.edu/research/vocal/default.html)a](https://www.doe.mass.edu/research/vocal/default.html), [DESE, 2023a](https://www.doe.mass.edu/research/vocal/default.html)) and is designed to show that the addition of 20 new items designed to measure student perceptions of [deeper learning](https://www.doe.mass.edu/deeperlearning/) did not alter the psychometric properties of the VOCAL measurement scale. Deeper learning (DL) for all students is an important goal of the Commissioner’s, [Our Way Forward](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.doe.mass.edu%2Fbese%2Fdocs%2Ffy2019%2F2019-06%2Fitem2.docx&wdOrigin=BROWSELINK) strategy report (DESE, 2023b). The data shown here are from the joint calibration as scaled scores for all students are derived from the joint calibration of students’ responses and items from all four participating grades.

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 students’ perceptions of school climate. All six validity aspects are addressed in this study; reporting of consequential validity is not covered in the same depth.

1. **School climate construct and raw response scoring**

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 (and nine topics) of school climate: engagement (cultural competence, relationships, and participation), safety (emotional, physical, and bullying), and environment (instructional, mental-health, and discipline). The conceptual framework and construct definitions are outlined in Table 1.

Table 1

VOCAL’s conceptual framework**1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Engagement** | **Safety** | **Environment** | |
| * **Cultural Competence:** The extent students feel adults/students value diversity, manage dynamics of differences, and avoid stereotypes. * **Relationships:** The extent students feel there is a social connection and respect between staff/teachers and students, and between students and their peers. * **Participation:** 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. | * **Emotional:** The extent students feel a bond to the school, and the extent adults/students support the emotional needs of students. * **Physical:** The extent that students feel physically safe within the school environment. * **Bullying:** 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. | | * **Instructional:** The extent that students feel the learning environment is collaborative, relevant, challenging, and supportive. * **Mental-health:** The extent that students have access to support systems that effectively support their social, emotional, and mental health well-being. * **Discipline:** The extent that discipline is fair, applied consistently and evenly, and a shared responsibility among staff, teachers, and students. |

A Likert scale with four response options was used to rate grade 4 students’ perceptions of school climate. Coding for all positively valenced items dictated that a response of “0” (*untrue*) indicated the least favorable views of school climate, with a “3” (*always* *true*) denoting the most favorable school climate. Response scoring categories “1” and “2” corresponded to *mostly untrue and mostly true*, respectively. Note, seventeen items across the four grade-level surveys were reverse scored: the majority of these items measure two topic areas of bullying and physical safety; ten bullying behavior items and five physical items are reversed scored. A higher item score, irrespective of whether the item is positively or negatively valenced, is associated with a more favorable school climate.

1. **2022 survey improvements**

The twenty new DL items were mostly developed for the participation topic that is related to the engagement dimension (10 items), and for the instructional environment topic of the environment dimension (7 items). Two other engagement items related to the cultural competence and relationships topics of engagement were developed, respectively. Lastly, one item related to emotional safety topic was developed for the two younger grades. The items prompts are shown in Table 2. These items were designed to capture whether deeper learning tasks were engaging students in the classroom and whether the conditions for deeper learning were embedded in teachers’ instructional strategies within their classroom environments.

The new survey specification for each of the four grade-level surveys is found in Appendix A (p.30). As a result of adding deeper learning items, students in grades 4 and 5 respond to a total of 46 items with students in grades 8 and 10 responding to a total of 48 items. The addition of the DL items to the participation topic and to the instructional environment topic allowed DESE to report out scores to schools for these two topic areas; schools already received a bullying topic score (safety dimension). For schools and districts to receive topic scores in 2022, they needed a minimum of 50 students responding to the respective topics. The number of items in these three topic areas represents 57%, 50%, and 57% of the items used to measure their three related dimensions (Appendix A). Across the four grade-level surveys, there are a total of 113 items that are used in the joint calibration to derive the scaled scores for each of the three dimensions, three topic areas, and for the overall school climate score.

Table 2: New deeper learning (DL) items added to 2022 VOCAL surveys

|  |  |  |
| --- | --- | --- |
| **ENGAGEMENT ITEMS** | | |
| Item code1 | Grade (s) | Item prompt |
| ENGCLC10 | 8 | In my academic classes, I work with groups of students who are from different backgrounds (for example, different races, cultures, family incomes, religions, sexes, or sexual orientations). |
| ENGREL15 | 4, 5 | In my classes, students work well together in groups. |
| ENGPAR15 | 4,5 | In my classes, students teach each other how they solved a problem. |
| ENGPAR16 | 5 | Students plan and work on group projects that solve real-world (everyday) problems. |
| ENGPAR17 | 8 | In at least two of my academic classes, students plan and work on projects that solve real-world problems. |
| ENGPAR18 | 8, 10 | In my academic classes, students review each other’s work and provide advice on how to improve it. |
| ENGPAR19 | 8 | In my classes, teachers use open-ended questions that make students think of many possible answers. |
| ENGPAR20 | 8 | I can connect what I learn in one class to what I learn in other classes. |
| ENGPAR21 | 8, 10 | In my academic classes, students wrestle with problems that don't have an obvious answer. |
| ENGPAR22 | 10 | In my academic classes, I am asked to apply what I know to new types of complex tasks or problems. |
| ENGPAR24 | 10 | In my school, students work on long-term group projects (more than one month in length) that they independently carry out. |
| ENGPAR25 | 4 | Students plan and work on group projects that solve real problems. |
| **ENVIRONMENT ITEMS** | | |
| Item code1 | Grade (s) | Item prompt |
| ENVINS17 | 4,5 | In my classes, it is OK for me to suggest other ways to do my work. |
| ENVINS18 | 5 | Teachers go over my work with me so I can improve it before it is graded. |
| ENVINS19 | 10 | In my school, teachers focus on my understanding of the material and not on my grades. |
| ENVINS20 | 10 | In my academic classes, there is a good balance between students having to master subject content and being able to explore topics that interest them. |
| ENVINS21 | 10 | In my classes, mistakes or even failure on an assignment are viewed as an important part of our learning. |
| ENVINS22 | 8 | Students are given multiple opportunities to show that they have mastered their classwork. |
| ENVINS23 | 4 | Teachers go over my work with me so I can improve it. |
| **SAFETY ITEMS** | | |
| Item code1 | Grade (s) | Item prompt |
| SAFEMO13 | 4,5 | I feel safe sharing my feelings in class. |

1ENG: Engagement; CLC: Cultural Competence; REL: Relationships; PAR: Participation; ENV: Environment; INS: Instructional environment;.SAF: Safety; EMO: Emotional safety

1. **Profile of 2022 respondents**

The sampling frame included students in grades 4, 5, 8 and 10; the profile of students in the sample in comparison to their representation in the state is shown in Table 3. 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.

Table 3.

Participating students’ profile1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Subgroup**  **(percent)** | **G4**  **sample** | **G4**  **state** | **G5**  **sample** | **G5**  **state** | **G8**  **sample** | **G8**  **state** | **G10**  **sample** | **G10**  **state** |
| Number of students | 50,932 | 67,417 | 52,878 | 68,508 | 56,796 | 73,044 | 37,562 | 73,238 |
| Percent response | 76% | 100% | 77% | 100% | 78% | 100% | 51% | 100% |
| Usable surveys | 50,932 | --- | 52,878 | --- | 56,796 | --- | 37,562 | --- |
| Percent usable | 76% | --- | 77% | --- | 78% | --- | 51% | --- |
| Fully complete surveys | 50,455 | --- | 52,457 | --- | 55,413 | --- | 36,369 | --- |
| Percent complete2 | 99.1% | --- | 99.2% | --- | 97.6% | --- | 96.8% | --- |
| Female | 49.2% | 48.6% | 48.7% | 48.5% | 48.6% | 48.4% | 48.4% | 48.3% |
| Male | 50.7% | 51.4% | 51.2% | 54.1% | 51.3% | 51.4% | 51.3% | 51.3% |
| Non-binary | <.01% | <.01% | <.01% | <.01% | 0.1% | 0.2% | 0.3% | 0.3% |
| Asian | 7.3% | 7.6% | 7.0% | 7.3% | 6.9% | 7.2% | 5.7% | 6.8% |
| Black | 8.7% | 9.3% | 8.9% | 9.3% | 9.0% | 9.5% | 9.0% | 9.4% |
| Hispanic | 23.9% | 23.9% | 23.7% | 23.8% | 22.8% | 23.4% | 21.5% | 22.8% |
| Other3 | 4.8% | 5.0% | 4.7% | 4.8% | 4.0% | 4.2% | 3.9% | 3.9% |
| White | 55.3% | 54.2% | 55.6% | 54.8% | 57.2% | 55.6% | 59.9% | 56.8% |
| Students with disabilities | 19.7% | 21.7% | 19.7% | 21.6% | 18.1% | 20.3% | 16.8% | 19.4% |
| English learners | 13.8% | 14.6% | 10.7% | 11.5% | 7.9% | 8.4% | 6.9% | 8.3% |
| Low Income | 47.0% | 47.4% | 46.3% | 47.0% | 45.0% | 46.6% | 42.2% | 43.9% |

1State numbers exclude students who took the MCAS-Alt test; 2Percent of students who provided a response to all items on the survey/number of usable surveys; 3Includes Multi-race, Non-Hispanic, Native American, and Native Hawaiian, Pacific Islander students

Response data indicated that over three-quarters of students in grades 4, 5, and 8 participated in the survey; these response rates are close to pre-pandemic levels. Only half of grade 10 students participated; their participation has not returned to pre-pandemic levels (it was 71% in 2019). The profile of the sample is reasonably representative of the state for each grade (Table 3). Students of color, low income, and English learner are all slightly under-represented in each grade. Students with disabilities are under-represented by 2% or more points across the four grades.

Roughly 97% or more of students fully completed their surveys. No surveys were excluded due to non-response of items. The Rasch model (Rasch, 1960) 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.

**5. Data analyses procedures**

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 (2019a) was used to perform Rating Scale model analyses of the data (Andrich, 1978a, 1978b). Technical details explaining the Rasch model are provided in the [2018 validity study](https://www.doe.mass.edu/research/vocal/default.html). 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. Items from all four grade-level surveys were concurrently calibrated. The common items across the four grades place the items on the same scale metric and then they are anchored to the 2018 scale. The results of the Rasch-based reliability and validity assessment for the 2022 VOCAL survey are summarized in the next section; when appropriate, the effect of adding the 20 DL items is discussed.

1. **Validity evidence for 2022 VOCAL survey (DL item addendum)**
   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 4 outlines the specific validity aspects addressed in this technical report. Because all students’ responses from the four grades are calibrated together, this report discusses results from the conjoint calibration, and, when appropriate, highlights the impact of adding the 20 DL items to the conjoint calibration.

The six aspects of construct validity (content, structural, substantive, generalizability, external, and consequential) are discussed in turn. *All validity evidence are shown in appendices and only a brief summary of the results are provided in the main body of the report.* *Please see prior reports for an explanation of each aspect of validity (*[*DESE, 2019a*](https://www.doe.mass.edu/research/vocal/default.html)*;* [*DESE, 2023a*](https://www.doe.mass.edu/research/vocal/default.html)*).* An overall school climate VOCAL score, an engagement dimension score, a safety dimension score, an environment dimension score, and a participation, bullying, and instructional environment climate topic score are reported out to districts and schools. The goal of these analyses is to ensure that DESE can report seven reliable and construct-relevant scores to schools and districts who serve the four grades. In addition, DESE examined whether the 20 new DL items altered the psychometric properties of the VOCAL scales. Appendix B (p. 31) provides a guide to the validity criteria used in this study for each aspect of construct validity.

Table 4

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.

* 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 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 C1 (113-item school climate scale; p.32), and in Appendix C2 (42-item engagement scale, p.33), Appendix C3 (34-item safety scale, p.34), Appendix C4 (37-item environment scale, p.35); Appendix C5 (24-item participation topic scale, p.36); Appendix C6 (17-item bullying scale, p.36), and Appendix C7(24-item instructional environment scale, p.37).

Using item fit evaluation criteria (Appendix B, p.31) to assess the overall 113-item VOCAL scale, 9 safety items (6 bullying items, 3 physical safety items) are considered degrading to measurement (means square error >2.0): all these safety items are reverse-scored. These results are similar to those found in prior validity studies ([DESE, 2019a](https://www.doe.mass.edu/research/vocal/default.html), [DESE, 2023a](https://www.doe.mass.edu/research/vocal/default.html)) and are explained by the reverse-coded items forming a “scoring method” factor (Conrad, Wright, McKnight, McFall, Fontana, and Rosenbeck, 2004). When these items are analyzed as part of the safety scale or bullying scale, these items are productive for measurement. These items are important to the construct relevance of the school climate scale, but particularly the safety and bullying scales, and are kept in the survey as a result. With the exception of these nine items in the overall VOCAL scale, items comprising the seven scales that are reported out are productive for measurement and of good technical quality.

An examination of the point-to-measure (PTM) correlations also indicates that the 113 VOCAL items are construct relevant. With the exception of ENVINS10 (My classwork is hard but not too hard), all PTMs range between 0.3 and 0.7 and thus meet the validity evaluation criteria for item technical quality outlined in Appendix B. When ENVINS10 is calibrated as part of the 21-item instructional environment scale, its PTM is 0.36 indicating it is construct relevant. **The DL items** which are highlighted in yellow in the relevant appendices all are well-fitting and contribute to productive measurement; the PTMs for these items are all above 0.3 indicating that for each scale produced, they are construct relevant.

**Content validity conclusion.** The fit analyses support the use of the scores at the dimension level and topic level (participation, bullying, instructional environment) and showed that the addition of the 20 DL items in the joint calibration did not disrupt the measurement system.

* 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, 2019a; Linacre, 2019b), (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 113 school climate items and separately for items belonging to each of the three dimensions and topics. The evaluation criteria used to assess structural validity are shown in Appendix B (p.31).

**Overall scale.** The variance explained is less than ideal of 40% (36.5%) for the overall school climate construct measure; this measure extracts all the common variance across all 113 items in the survey (Appendix D1, p.38). 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 6.3 times the variance explained by the first contrast (residual), meeting Linacre’s criterion of at least 4 times (Linacre, 2019b).

**Dimension scales.** The 42-item engagement scale and 37-item environment scale meet the evaluation criteria and support the structural validity of the two scales (Appendix D2, p.38). The 34-item safety scale explains almost 40% of the variance explained with the first contrast’s residual variancemeeting the criteria ofless than 5%. However, the variance explained by the safety dimension is only 2.7 times the variance explained by the first contrast. Similar to prior years, the positively valenced bullying protective items are separating from the bullying behavior reverse-coded items; in the [2018 validity study](https://www.doe.mass.edu/research/vocal/default.html), an empirical analyses showed that the “signal-to-noise of this separation was not of sufficient magnitude to distort measures (p.31-33)” for the safety dimension.

**Topic scales.** The variance explained by each of the three topic scales (participation, bullying, and instructional environment) was over 40% (Appendix D3, p.38). The first contrast’s residual variance was less than 5% of the total observed variance for the participation and instructional environment topics. Similarly, the variance explained by these topic scales was near to or at 4.0 times the variance explained by the first. These data support the structural validity of the participation and instructional topic scales. Similar to prior years, the reverse coded bullying behavior items separated from the positively valanced bullying protective items leading to the unexplained variance of the 1st contrast being over 5% and a 1st contrast multiple of less than 4. Both bullying protective and bullying behavior items are essential to measuring the bullying climate in schools and the correlation for the two sets of items was close to 0.7 indicating they are related constructs (data not shown). All bullying items were retained.

**Dimension and topic score 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 the three separately calibrated topics (participation, bullying, and instructional environment). All correlations between the three dimensions are between 0.5 and 0.9 (Appendix E, p.39). Not unexpectedly, the correlation is highest between students’ engagement dimension scores and students’ environment dimension scores (0.85); these two dimensions are heavily influenced by the participation and instructional environment topics that make up over 50% of the items in these dimensions (Appendix A, p.30). Participation examines student perceptions of whether they feel engaged intellectually, behavioral, and emotionally in the classroom with the instructional environment items targeted to assessing students’ views on whether teachers set up the conditions for learning needed to engage students in the classroom. These closely related topics drive the strong correlation between the engagement and environment dimensions.

The correlations between the three topics and the three dimensions are as expected with the strongest correlations between students’ topic scores and their associated dimension scores (for example, students’ bullying scores have a correlation of 0.92 with their associated safety dimension scores; they have a weaker correlation with students’ engagement (0.59) and environment (0.60) scores, respectively). Among the three topic areas, the strongest correlation is between students’ participation scores and instructional environment scores (0.75); given the two topics measure different aspects of the classroom climate, this finding is predictable.

**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 (113-items) and shows that the addition of 20 DL items did not negatively affect the structural validity of the VOCAL scale. 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 dimensionality and residual analyses of all the sub-scales also support their use for their intended purpose (to provide more granular data to schools and districts so they can identify strengths and areas for growth). The residual analyses highlighted that bullying behavior items separated from the bullying prevention items; however, the signal-to-noise of this separation is not of sufficient magnitude to distort measures. The correlations of the sub-scale dimension and topic scores (all students) support the theoretical premise that the school climate construct is composed of three related, but distinct dimensions of school climate and the school climate topics contribute to the content and structural validity of the school climate instrument.

* 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 (Wright & Stone, 1979) and meets survey developers’ *a priori* expectations.

**Rating scale.** The rating scale for the 113 items of the VOCAL survey functioned well (Appendix F, p.40). Except for the little used score category of zero (for most items, response of “never true”), the category threshold fit statistics are excellent with MNSQ error near to 1.00 and do not differ from those reported in the [2018 calibration](https://www.doe.mass.edu/research/vocal/default.html) using the three older grades. 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. The addition of the 20 new DL items did not perturb the rating scale and the rating scale is being used as intended by these students.

**Overall VOCAL item hierarchy.** The ordered pattern of item difficulties continues to conform to best test design principles (Wright and Stone, 1979). Appendix G1 (p. 41) and Appendix G2 (p. 42) display the item-variable maps for the 113-item VOCAL survey. In Appendix G1, the engagement items are shown in yellow, safety items in pink, and environment items shown in green. In Appendix G2, the dimension items are color coded by topic area. Items for each dimension and topic span the breadth of the school climate continuum with items from different dimensions/topics overlapping as you move from low to high on the scale metric. Appendix G3 (p. 43) shows the item threshold-variable map; calibrations cover 99.4% 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.

**Substantive validity conclusion.** The well-functioning rating scale combined with the theoretically grounded 113-item item hierarchy provides the evidence needed to support the substantive validity aspect of the school climate construct.

* 1. **Generalizability**

A measure is considered generalizable when the score meaning and properties function similarly across multiple contexts (e.g., grades, student groups, 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 also used Differential Item Functioning (DIF) analyses to empirically test for item invariance across several student groups; item invariance ensures comparability of score interpretation.

**VOCAL scale reliability evidence (113-item, all 4 grades).** Best test design principles (Wright & Stone, 1979) necessitate the alignment of the mean of the item distribution to the mean of the person distribution. The mean person difficulty of the 113-item scale was +1.14 logits with a standard deviation of 0.99 logits (Appendix H1, p.44). The items are reasonably well targeted for the student distribution resulting in a real person separation reliability (PSR) of 0.93. Notable in the item-variable map (Appendix G1, p. 41) is the relative rarity of some bullying and aggressive behaviors when compared to other indicators assessed; these off-target items likely contribute to the misalignment of the person and item distributions.

A real person separation index (PSI) of greater than 3 indicates that the students can be reliably divided into 4 scoring groups. The overall scale has a real PSI of 3.58 indicating that the student distribution based on the overall VOCAL scale can reliably be divided into 5.1 statistically distinct groups. This provides evidence for the reproducibility and stability of the school climate construct even with the 20 new DL items added to the calibration. It also provides evidence needed to reliably divide students’ scores into four “favorability” categories; a new report was made available to schools and districts that provided them with the percent of students in each of the following four categories: “ least favorable” (average score <=30), “somewhat favorable” (31 to 50), “favorable” (51 to 70), and “most favorable” (>70). The PSI indicates that these scoring categories are statistically distinct.

**Grade-level reliability evidence.** When calibrated separately, the real PSR for each of the four grade-level forms was 0.90 or greater with PSIs of greater than three; as a result, the variance of the four grades’ score distributions is sufficient to allow scores to be divided into four reliable scoring categories (Appendix H1, p.44). The reliability data supports the generalizability of the four grade-level survey scores if they were to be calibrated separately.

**Dimension/Topic reliability evidence.** Appendix H2 (p. 45) and Appendix H3 (p. 45) shows the reliability of the three dimensions and three topics, respectively. The new items added to the 2022 maintained or improved the reliability of all scales. The real PSR of the engagement, safety, and environment scores was 0.84 (0.80 in 2019), 0.83 (0.81 in 2019), and 0.82 (0.81 in 2019), respectively. These reliabilities are also likely attenuated due to the design of the test forms (Schwartz, Ayers, and Wilson, 2017). Students across the four grades only respond to a small sub-set of common items for each dimension; students largely respond to sets of unique items. As a result, a large amount of “missing data” is realized when the four 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).

With fewer items, the reliability of the three topic scales met the 0.7 minimum reliability target; the real PSR for participation, bullying, and instructional environment was 0.74, 0.73, and 0.71, respectively. These reliabilities are similarly disattenuated as students respond to many unique items for each topic (Schwartz, Ayers, and Wilson, 2017). Schools and districts receive a topic score if there are 50 or more students contributing to the score; this minimum threshold determined by sensitivity analyses (data not shown) was designed to help ensure that schools and districts receive reliable data for the three topics.

**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 student groups of respondents (e.g., gender, race/ethnicity). Respondents with the same ability (endorsement level), should have the same probability of affirming an item irrespective of the student group they belong to. Data are shown in Appendix I1 (gender, p. 46), Appendix I2 (race/ethnicity, p. 47), Appendix I3 (low income, p. 49), Appendix I4 (English learner, p. 50), and Appendix I5 (students with disabilities, p. 51), respectively. In this study, items in each appendix were flagged if they average difficulties differed by 0.5 logits (flagged by coloring cell yellow); items exhibiting severe DIF (>=0.70) were, in addition, highlighted by bolding and coloring the text brown. Due to the addition of the DL items, schools and districts would now receive dimension scores broken out by student group (minimum N of 20 students). As a result, the DIF analyses were also performed at the dimension level and the pertinent DIF size results are shown in the last column of each appendix.

For the VOCAL scale as a whole and for each respective dimension calibration, the analyses indicated that item thetas did not differ significantly across the following student groups: gender, race/ethnicity, low income, and students with disabilities. For the race/ethnicity and student with disabilities student group comparisons, one item exhibited severe DIF. With only one item exhibiting DIF and given the length of the scales (VOCAL-, dimension-level), one item was not expected to lead to construct-irrelevance variance and perturb the measurement system.

**English learner DIF.** For the conjointly calibrated VOCAL scale, severe DIF was present for English learners with thirteen items having DIF of greater than 0.70 logits (Appendix I4). Eight of the thirteen items (ENGPAR3, ENGPAR14, SAFEMO6, SAFBUL10, SAFBUL11, SAFBUL16, SAFPSF5, ENVINS12) were on the grade 10 form. Five of these eight items (ENGPAR3, ENGPAR14, SAFBUL10, SAFBUL11, SAFPSF5) still exhibited DIF when the calibration was run for each of the dimensions separately (Appendix I4, p. 50). Two of the thirteen DIF items (ENGPAR1, SAFBUL12) were administered on the grade 4 and 5 forms; ENGPAR1 still exhibited severe DIF in the engagement dimension calibration. Three items, SAFPSF4, SAFBUL5 and SAFBUL15, exhibited DIF on the grade 8 VOCAL form; SAFPSF4 still exhibited DIF in the safety dimension calibration. In terms of valence and content, eight of the thirteen DIF items required reverse scoring; nine of the thirteen items were from the safety dimension, three from the participation topic within the engagement dimension, and one from the instructional environment topic within the environment dimension. DESE’s surveys were not translated for English learners. Language barriers likely explain the DIF present across items for EL students; EL students are unable to fully access the survey content.

The EL results from the 2022 survey largely replicate the DIF results from the [2018 study](https://www.doe.mass.edu/research/vocal/default.html) and from the [grade 4 addendum validity study](https://www.doe.mass.edu/research/vocal/default.html) (grade 4 was added in 2019). In 2018, analyses were performed to determine the impact of including the DIF items in the EL students’ overall school climate score estimations. EL students’ overall school climate scores were estimated with and without these DIF items included. As a result of these analyses, the decision was made to remove the DIF items from the overall school climate calibration process in 2018 when estimating EL students’ scores *for grade 10 students*; this process was repeated for the calibration of EL students in 2019 and now again in 2022. In 2022, when estimating dimension scores for EL students in grade 10, the DIF items were also removed when the dimension scales were calibrated separately. As a result, the overall and dimension scores for grade 10 EL students are based on a smaller number of items; the resulting overall scale has a PSR reliability of over 0.9 for EL students and a safety dimension reliability of over 0.8 (data not shown). This helps ensure that high schools and districts receive reliable and unbiased aggregate scores for their grade 10 EL students.

**DIF analyses of the new DL items.** Of the 20 DL items added to the VOCAL scale,one item, ENGPAR24, exhibited moderate DIF (0.69 logits) when comparing EL students to not-EL students. If the EL DIF for ENGPAR24 (administered in grade 10 only) replicates in future years, this item should be removed from the EL calibration for both the overall VOCAL calibration and the engagement dimension calibration. The remaining nineteen DL items functioned well across the five student group comparisons and will provide reliable measures for student group analyses.

**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 student group comparisons. *2022 scores for grade 10 EL students should be viewed with caution due to the decreased number of items used to estimate EL student group scores. However, the scale reliabilities at the VOCAL scale level and at the dimension level are good, indicating that they provide a reliable picture of EL students views of school climate.* The addition of the 20 DL items did not perturb the measurement system in terms of DIF across all items administered with only one item exhibiting DIF when EL students were compared to not-EL students. This indicates that student groups likely had similar interpretations of the DL item statements which will continue to allow student group scores to be reliably analyzed for differences.

* 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.

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: attendance, chronic absence, discipline rates, suspension rates, and academic achievement and growth.

**Student-level responsiveness.** The responsiveness of an instrument is measured by the person strata index, H, which provides the number of statistically distinct scoring 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 3.58, the number of person strata for the 113-item VOCAL instrument is equivalent to 5.1 distinct person strata (Appendix H1, p. 41). The addition of 20 new DL items improved the responsiveness of the overall instrument slightly from 4.8 in 2019 ([DESE, 2023a](https://www.doe.mass.edu/research/vocal/default.html)) to 5.1 in 2022. 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) when all four grades are conjointly calibrated, and this level of responsiveness supports dividing students’ scores into four “favorability” categories.

**Concurrent Validity: Behavioral indicators.** Preliminary evidence of concurrent validity at the school level indicates a significant relationship between students’ overall school climate scaled scores and several behavioral criterions. All the statistically significant associations were in the expected direction; these data are summarized in Appendix J1 (p. 52).

The results reveal that there are small to moderate, statistically significant, relationships between the overall VOCAL scaled-scores and attendance (0.21), chronic absence rates (‑0.23), in-school suspension rates (-0.29), and out-of-school suspension rates (‑0.39). The magnitude of the correlation between the overall VOCAL scaled scores and disciplinary rates is considered large (‑0.58). When the relationship was examined between schools’ dimension scores and the five behavioral indicators, the strongest correlations appear between the safety dimension scores and the five behavior indicators (less so for the in-school suspension indicator). These patterns and magnitude of associations, largely mirror those found in prior validity studies ([DESE, 2019a](https://www.doe.mass.edu/research/vocal/default.html), [DESE, 2023a](https://www.doe.mass.edu/research/vocal/default.html)); replication of results helps add to the validity argument supporting the external validity of the VOCAL scale(s).

**Concurrent Validity: Academic indicators.** The relationship between academic indicators and the overall VOCAL scores of schools was broken out by grade (Appendix J2, p.52). The results indicate that there are small to moderate, statistically significant, relationships between schools’ overall VOCAL scaled-scores and English Language Arts and Literacy (ELA) scaled scores (e.g., 0.30 in grade 4), ELA growth scores (e.g., 0.22 in grade 5), Mathematics scaled scores (e.g., 0.26 in grade 8), and between schools’ mathematics growth scores (e.g., 0.18 in grade 8). For the four grades, the patterns and magnitude of the correlations replicate correlations determined in 2019 ([DESE 2023a](https://www.doe.mass.edu/research/vocal/default.html)); similar to 2019, in 2022, there was not a significant relationship between high schools’ academic indicators and the overall VOCAL score. The 2022 findings somewhat replicate the findings from 2018 for grade 10; in 2018, the correlation between high schools’ achievement scores was statistically significant (small in size) but not so for students’ growth scores ([DESE, 2019a](https://www.doe.mass.edu/research/vocal/default.html)). This difference between 2019 or 2022 and 2018 may be explained by the move from DESE’s old legacy MCAS tests in 2018 to the new generation MCAS tests that were administered in 2019 and in 2022.

When the relationship was examined between schools’ dimension scores and the four academic indicators within each grade level, the strongest correlations appear between the safety dimension scores and the four academic indicators across all grades (Appendix J3 to J6, p. 53). For example, the correlation between schools’ safety scores and ELA scaled scores is 0.47 in grade 4; this compares to 0.04 and 0.14 for engagement and environment dimension correlations, respectively. In grade 10, safety scores have a small to moderate positive relationship with the four academic indicators and the associations are significant; however, there is no apparent relationship in grade 10 between engagement scores and the four academic indicators or between environment scores and growth scores. In grade 10, the environment scores are negatively related to schools’ achievement scaled scores, indicating students in schools with more favorable views of the school environment have, on average, lower achievement scores (opposite to the expected direction). The grade 10 MCAS test is a competency assessment (high stakes), and this may help explain the attenuated correlations in this grade. *Further research is needed to help explain why the direction of some of the correlations is opposite to expectations for this grade.* Irrespective of the grade, safety seems to be a key driver of the positive and significant relationships between the four academic indicators and the overall VOCAL score.

**External validity conclusion.** Overall, the external validity evidence supports the conclusion that the school climate surveys are responsive and can measure change in student perceptions of school climate. The pattern of correlations (direction and magnitude) provides evidence to support VOCAL’s external validity with the five behavioral indicators; replication of the associations between these behavioral indicators and VOCAL scores across years strengthens the external validity argument. Similarly, with the exception of grade 10, the pattern of correlations (direction and magnitude) indicate that overall VOCAL scores (and most dimension scores) have a positive statistically significant relationships with the four academic indicators. Evidence for grade 10 is less conclusive as the relationship between the four academic indicator scores and VOCAL scores is either non-significant or of diminished magnitude (and in terms of growth scores, the direction of the relationship is opposite to what was expected for the environment dimension score). These attenuated correlations replicate prior years for grade 10 and may indicate the high stakes associated with the MCAS test in grade 10 is impacting the relationships between students’ academic achievement/growth and school climate perceptions. It is important to restate that, irrespective of the grade, safety seems to be a key driver of the positive and significant relationship between the four academic indicators and the overall VOCAL score.

Readers should be aware that correlational cross-sectional data do not support the interpretation that more positive school climates lead to (cause) improved student academic achievement or growth, or to improved behavioral outcomes. In addition, these simple correlations do not account for the nested nature of educational data. Future validity work should 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 (DESE, 2019b, 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 students contributing to the overall school climate score, and an N of 20 students for dimension scores. 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, the potential consequences attached to the use of the scores is not zero.

**Inclusion of deeper learning items**. The inclusion of 20 DL items could have negative consequences and affect the validity of the VOCAL data; their inclusion could have perturbed the measurement system by resulting in student survey fatigue. The addition of the DL items increased the length of the survey in all four grades. The total number of items in grades 4 and 5 increased from 40 in 2019 to 46 in 2022; similarly, the total number of items in grades 8 and 10 increased from 40 in 2019 to 48 in 2022. To determine if survey fatigue was an issue, the completion rates from 2022 were compared to those from 2019. The results are shown in Table 5.

Table 5

Comparison of survey completion rates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | Grade 4 | Grade 5 | Grade 8 | Grade 10 |
| 2019 | 99.3% | 99.5% | 97.3% | 97.5% |
| 2022 | 99.1% | 99.2% | 97.6% | 96.8% |

1Percent of students who provided a response to all items/number of usable surveys

The difference in survey completion rates between 2019 and 2022 was less than 1% across all four grades. Students in 2022 appear to have coped with the lengthened surveys with almost 97% or more students responding to all items on each grade-level survey. The benefit of the added items was that it allowed the reporting of two additional topic scores, namely, the participation topic (engagement dimension) and the instructional environment (environment). These two topics areas are targeted at the classroom climate within schools and provide scores that can be monitored over time as more schools and districts implement deeper learning strategies in the classroom. The benefit of the longer surveys was deemed to outweigh the cost of potential student fatigue.

**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, 2019c). DESE has made progress in making the survey more aligned to DESE’s deeper learning goals (DESE, 2023b). However, the lack of translating the survey into other languages undoubtedly led to English learner students (particularly grade 10 EL students) not having full access to survey content. 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.

**7.0. VOCAL report conclusion**

This validity study was designed to provide psychometric evidence to justify the use of 2022 VOCAL scores by schools and districts within Massachusetts who serve grades 4, 5, 8, and 10. New to the 2022 VOCAL administration, was the addition of 20 deeper learning items. It was important to establish that the addition of these items did not perturb the measurement system given that all students and all items from all four grades would be conjointly calibrated to produce the VOCAL scores for districts and schools. When conjointly calibrated to obtain an overall school climate score, there are a total of 113 items administered across the four grades. Evidence was provided that supported each aspect of construct validity (content, structural, substantive, generalizability, external, and consequential) for the 113-item school climate measure and any impact of including the DL items was discussed.

A large majority of the 113-item VOCAL survey fit the Rasch model well and were of good technical quality; a “scoring method” factor made up of 11 reverse-scored items misfit the model with 8 of the 11 potentially degrading the measure. When calibrated separately by dimension or topic, none of these 8 items’ misfit was at the level to degrade the measures. Despite these misfitting items, the VOCAL scale met the unidimensional assumption of the Rasch model. Students’ dimension scores (engagement, safety, and environment) and topic scores (participation, bullying, and instructional 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 and of the new DL items met developers’ *a priori* expectations. This summarization supports the content, structural, and substantive validity of the VOCAL survey.

The VOCAL survey was reliable at the school-, grade-, and dimension/topic- level. With the exception of English learners, differential item function analyses indicated that students from different student groups with the same score had, within measurement error, the same affirmation level and likely interpretation of each VOCAL item. The new DL items performed well with only one item exhibiting DIF for EL students (grade 10 item). These data support the generalizability of the school climate construct with the new DL items included.

Schools were meaningfully differentiated by their school climate scores due to the responsiveness of the VOCAL scales. VOCAL scores were, as expected, appropriately related (positively or negatively) to school-level behavioral indicators and, with the exception of grade 10, positively related to students’ school level achievement and academic growth; this digest supports the external validity of the VOCAL survey.

The replication of each of the validity analyses across grades and years provides further evidence to support the validity argument that the VOCAL survey is providing reliable reproducible scores and can be used for its intended purpose. In conclusion, the psychometric properties of the 2022 VOCAL instrument with the new DL items met the assumptions of the Rasch-model, namely the items are well-fitting, invariant, and form a unidimensional scale. The VOCAL reports distributed to schools and districts provide theneeded data to inform and support continuous school climate improvement.

**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.

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.

DESE (2019a). [2018 Views of Climate and Learning (VOCAL) Validity Study](https://www.doe.mass.edu/research/vocal/2018/validity-study.docx). Available for download at <https://www.doe.mass.edu/research/vocal/default.html>

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

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

DESE (2023a). 2019 Views of Climate and Learning (VOCAL): Grade 4 validity addendum. Available for download at <https://www.doe.mass.edu/research/vocal/default.html>

DESE (2023b). Our way forward for Massachusetts K-12 public education: The Commissioner’s report to the Board. Available for download [here](https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.doe.mass.edu%2Fbese%2Fdocs%2Ffy2019%2F2019-06%2Fitem2.docx&wdOrigin=BROWSELINK).

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.

Linacre, J. M. (2019a). Winsteps® Rasch measurement computer program. Portland, Oregon: Winsteps.com

Linacre, J. M. (2019b). *A user’s guide to Winsteps, Ministep Rasch-model computer programs: program manual*, Chicago, US: MESA Press.

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

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).

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.

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

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

Appendix A: 2022 VOCAL Survey Specifcation by Grade

This picture shows the 2022 VOCAL survey specification by grade. Included is the number of items associated with each dimension/topic of school climate by grade (and in total). 

**1Common items removed from item tally**

**Appendix B: 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 thetas correspond to theoretical expectations. * Item/person variable maps. | Qualitative assessment of items in the construct and/or dimensions/topics. |
| **Generaliz-ability** | Item invariance and  Differential Item Functioning (DIF) | * Within standard error, items should retain same item difficulty (thetas) across administrations and survey forms (correlation of greater than or equal to 0.9). * For DIF, recommended criteria; In this study: <**0.5, good; 0.5 to 0.7, moderate DIF; >=0.7, severe DIF**) | 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 or topic-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 C1 (Content validity): Technical quality (mean-square error) of 113-item VOCAL scale**

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

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

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

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

**| 53 93105 37135 -1.37 .01|2.20 9.90|2.52 9.90|A .19 .40| 52.8 61.4| .00| SAFBUL11|**

**| 52 93503 37161 -1.40 .01|2.27 9.90|2.50 9.90|B .22 .40| 53.9 61.5| .00| SAFBUL10|**

**| 49 99575 50775 .82 .01|2.30 9.90|2.35 9.90|C .38 .54| 27.3 53.2| .00| SAFBUL7 |**

**| 47 120344 56638 -.38 .01|2.26 9.90|2.33 9.90|D .28 .46| 30.7 55.0| .00| SAFBUL5 |**

**| 75 191861 103727 .93 .00|2.20 9.90|2.28 9.90|E .33 .55| 29.0 51.9| .00| SAFPSF7 |**

**| 44 101714 52747 .66 .01|2.17 9.90|2.23 9.90|F .39 .55| 29.4 52.7| .00| SAFBUL2 |**

**| 54 219221 103536 .36 .00|1.93 9.90|2.08 9.90|G .29 .53| 41.1 54.6| .00| SAFBUL12|**

**| 73 139482 56459 -1.28 .01|1.94 9.90|2.07 9.90|H .29 .41| 52.3 59.6| .00| SAFPSF4 |**

**| 74 88163 37229 -.96 .01|1.81 9.90|1.99 9.90|I .31 .43| 47.8 57.6| .00| SAFPSF5 |**

**| 56 101616 56453 .29 .01|1.85 9.90|1.91 9.90|J .38 .49| 32.9 51.5| .00| SAFBUL14|**

**| 11 236056 103667 -.02 .00|1.77 9.90|1.78 9.90|K .34 .51| 47.6 57.0| .00| ENGPAR1 |**

**| 68 110325 93915 1.48A .00|1.51 9.90|1.78 9.90|L .27 .52| 39.2 45.0| -.07| SAFEMO11|**

**| 79 59076 50838 2.31 .01|1.59 9.90|1.69 9.90|M .36 .56| 38.4 46.5| .00| ENVDIS3 |**

**| 50 252531 103720 -.46 .01|1.51 9.90|1.61 9.90|N .40 .48| 58.9 60.8| .01| SAFBUL8 |**

**| 59 108398 50761 .44 .01|1.61 9.90|1.57 9.90|O .49 .52| 44.2 54.8| .01| SAFBUL17|**

**| 55 356828 197649 .61A .00|1.47 9.90|1.60 9.90|P .38 .55| 44.3 51.4| .04| SAFBUL13|**

**| 58 75273 37069 -.15 .01|1.42 9.90|1.57 9.90|Q .33 .48| 49.8 54.5| .00| SAFBUL16|**

**| 63 88242 37340 -.95 .01|1.57 9.90|1.53 9.90|R .36 .43| 51.6 57.5| .00| SAFEMO6 |**

**| 76 75927 37176 -.17 .01|1.43 9.90|1.54 9.90|S .34 .48| 49.1 54.6| .00| SAFPSF8 |**

**| 113 51861 37173 1.05 .01|1.42 9.90|1.54 9.90|T .40 .53| 39.8 46.1| .00| ENVMEN9 |**

**| 19 244602 103761 -.37A .01|1.44 9.90|1.52 9.90|U .29 .48| 50.8 59.9| .14| ENGPAR9 |**

**| 3 96681 56315 .44 .01|1.41 9.90|1.49 9.90|V .31 .50| 45.5 50.2| .00| ENGCLC3 |**

**| 4 76293 37239 -.19 .01|1.44 9.90|1.46 9.90|W .39 .48| 46.5 54.7| .00| ENGCLC4 |**

**| 24 101114 93496 1.57 .00|1.37 9.90|1.41 9.90|X .38 .52| 38.1 45.1| .00| ENGPAR14|**

**| 8 162941 103474 1.45 .00|1.33 9.90|1.39 9.90|Y .39 .57| 43.6 49.1| .00| ENGCLC8 |**

**| 71 256814 103766 -.86A .01|1.38 9.90|1.30 9.90|Z .51 .45| 64.6 65.8| .29| SAFPSF1 |**

**| 48 135945 50825 -1.19 .01|1.37 9.90|1.23 9.90| .46 .40| 76.2 72.1| .00| SAFBUL6 |**

**| 57 107712 56485 .08 .01|1.29 9.90|1.36 9.90| .33 .49| 50.9 52.9| .00| SAFBUL15|**

**| 97 147080 103545 1.73 .00|1.33 9.90|1.36 9.90| .49 .57| 42.7 47.7| .00| ENVINS14|**

**| 65 199746 103561 .70A .00|1.33 9.90|1.31 9.90| .55 .54| 47.3 53.2| .07| SAFEMO8 |**

**| 77 175088 146555 1.90A .00|1.28 9.90|1.33 9.90| .47 .57| 45.8 46.7| -.28| ENVDIS1 |**

**| 107 115243 52762 -.08A .01|1.33 9.90|1.27 9.90| .52 .51| 54.3 56.7| .18| ENVMEN1 |**

**| 6 251913 103512 -.38A .01|1.30 9.90|1.28 9.90| .40 .48| 58.2 60.0| -.07| ENGCLC6 |**

**| 31 159829 93295 .46 .00|1.10 9.90|1.27 9.90| .19 .50| 51.0 50.3| .00| ENGPAR21|**

**| 33 48825 37030 1.18 .01|1.21 9.90|1.27 9.90| .40 .53| 41.3 45.5| .00| ENGPAR24|**

**| 78 243990 103680 -.30A .01|1.27 9.90|1.23 9.90| .53 .49| 59.5 59.2| .08| ENVDIS2 |**

**| 93 204292 103781 .68 .00|1.13 9.90|1.26 9.90| .22 .54| 56.3 53.2| .00| ENVINS10|**

**| 70 160898 103496 1.49 .00|1.25 9.90|1.25 9.90| .59 .57| 43.6 49.0| .00| SAFEMO13|**

**| 62 145561 93536 .84A .00|1.23 9.90|1.24 9.90| .56 .51| 40.9 47.2| -.09| SAFEMO4 |**

**| 108 114384 50786 .16 .01|1.24 9.90|1.18 9.90| .51 .50| 53.9 56.2| .00| ENVMEN2 |**

**| 72 93778 56516 .55 .01|1.15 9.90|1.23 9.90| .38 .50| 49.2 49.6| .00| SAFPSF3 |**

**| 85 221700 103499 .30 .00|1.23 9.90|1.23 9.90| .53 .52| 53.0 55.0| .00| ENVDIS9 |**

**| 101 113827 52735 .15 .01|1.22 9.90|1.20 9.90| .49 .53| 51.8 55.4| .00| ENVINS18|**

**| BETTER FITTING NOT SHOWN +----------+----------+ | | | |**

**| 95 146900 93595 .79A .00| .82 -9.90| .83 -9.90|s .57 .51| 53.6 47.5| -.06| ENVINS12|**

**| 32 73180 37050 -.03 .01| .82 -9.90| .82 -9.90|r .53 .49| 62.5 54.0| .00| ENGPAR22|**

**| 42 216362 103516 .42 .00| .81 -9.90| .81 -9.90|q .50 .53| 62.8 54.5| .00| ENGREL15|**

**| 67 191197 109332 .63A .00| .81 -9.90| .81 -9.90|p .59 .55| 57.5 50.7| .05| SAFEMO10|**

**| 86 288857 146742 .22A .00| .79 -9.90| .81 -9.90|o .48 .53| 61.3 53.2| -.05| ENVINS1 |**

**| 94 445325 197206 -.13A .00| .81 -9.90| .79 -9.90|n .64 .52| 61.5 56.1| -.20| ENVINS11|**

**| 98 75920 37492 -.25A .01| .79 -9.90| .77 -9.90|m .60 .48| 62.3 54.9| .11| ENVINS15|**

**| 30 100713 56432 .32 .01| .77 -9.90| .78 -9.90|l .52 .50| 58.7 51.3| .00| ENGPAR20|**

**| 23 256367 160014 1.12 .00| .75 -9.90| .77 -9.90|k .56 .57| 58.9 49.1| .00| ENGPAR13|**

**| 39 203842 94162 -.47A .00| .75 -9.90| .76 -9.90|j .55 .46| 64.1 55.5| .01| ENGREL6 |**

**| 99 94013 52749 .95 .01| .76 -9.90| .76 -9.90|i .62 .56| 60.9 51.3| .00| ENVINS16|**

**| 103 61049 37028 .60 .01| .75 -9.90| .75 -9.90|h .60 .51| 57.4 49.2| .00| ENVINS20|**

**| 105 116505 56446 -.25 .01| .75 -9.90| .75 -9.90|g .59 .47| 62.5 54.4| .00| ENVINS22|**

**| 36 191077 103722 .94 .00| .72 -9.90| .73 -9.90|f .57 .56| 63.6 51.9| .00| ENGREL2 |**

**| 66 59221 37066 .77A .01| .71 -9.90| .72 -9.90|e .59 .52| 57.1 47.9| -.08| SAFEMO9 |**

**| 29 112956 56410 -.11 .01| .70 -9.90| .70 -9.90|d .52 .48| 65.3 54.0| .00| ENGPAR19|**

**| 38 196031 103516 .86A .00| .68 -9.90| .69 -9.90|c .56 .55| 65.6 52.4| -.02| ENGREL4 |**

**| 40 193156 93560 -.17A .00| .67 -9.90| .66 -9.90|b .64 .48| 65.2 54.4| -.07| ENGREL13|**

**| 35 152273 94117 .75A .00| .64 -9.90| .65 -9.90|a .55 .51| 60.2 47.8| -.11| ENGREL1 |**

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

**| MEAN 164936 81993 .12 .01|1.16 2.24|1.17 1.02| | 54.6 54.4| .00| |**

**| P.SD 89539.0 41586 .77 .00| .38 9.15| .42 9.41| | 9.6 5.4| .00| |**

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

**Appendix C2, (Content validity): Technical quality (mean-square error) of 42-item Engagement dimension scale1**

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

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

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

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

**| 11 236056 103667 -.03 .00|1.70 9.90|1.72 9.90|A .37 .49| 47.6 57.3| .00| ENGPAR1 |**

**| 19 244602 103761 -.37A .01|1.33 9.90|1.40 9.90|B .35 .47| 54.4 59.5| .14| ENGPAR9 |**

**| 3 96681 56315 .52 .01|1.30 9.90|1.35 9.90|C .39 .50| 47.5 50.7| .00| ENGCLC3 |**

**| 4 76293 37239 -.17 .01|1.31 9.90|1.32 9.90|D .45 .48| 48.3 55.2| .00| ENGCLC4 |**

**| 6 251913 103512 -.38A .01|1.23 9.90|1.20 9.90|E .42 .47| 59.9 59.9| -.08| ENGCLC6 |**

**| 24 101114 93496 1.64 .00|1.20 9.90|1.22 9.90|F .48 .53| 41.0 45.4| .00| ENGPAR14|**

**| 13 85845 37412 -.76 .01|1.19 9.90|1.17 9.90|G .47 .45| 55.4 56.8| .00| ENGPAR3 |**

**| 2 274655 103558 -1.19A .01|1.18 9.90| .99 -1.13|H .50 .41| 75.2 70.3| -.03| ENGCLC2 |**

**| 8 162941 103474 1.45 .00|1.15 9.90|1.18 9.90|I .49 .57| 47.7 49.1| .00| ENGCLC8 |**

**| 37 279115 103732 -1.27A .01|1.12 9.90| .99 -1.25|J .46 .40| 77.0 71.6| -.12| ENGREL3 |**

**| 21 67911 37467 .15A .01|1.09 9.90|1.11 9.90|K .51 .50| 50.4 53.5| .16| ENGPAR11|**

**| 12 124418 56162 -.59A .01|1.10 9.90|1.07 9.90|L .50 .46| 59.4 56.0| .07| ENGPAR2 |**

**| 31 159829 93295 .52 .00| .97 -6.73|1.08 9.90|M .33 .51| 54.7 50.8| .00| ENGPAR21|**

**| 33 48825 37030 1.21 .01|1.05 8.02|1.08 9.90|N .50 .54| 45.3 46.4| .00| ENGPAR24|**

**| 16 251838 103652 -.54A .01|1.07 9.90|1.00 .27|O .51 .46| 63.5 61.9| .10| ENGPAR6 |**

**| 17 207405 103532 .61A .00|1.06 9.90|1.01 2.52|P .61 .53| 53.3 54.4| .01| ENGPAR7 |**

**| 1 222161 93934 -.91A .01|1.05 9.90|1.02 3.31|Q .51 .44| 60.2 57.6| -.02| ENGCLC1 |**

**| 10 133580 56396 -.92 .01|1.05 8.45|1.03 3.96|R .45 .44| 59.5 57.3| .00| ENGCLC10|**

**| 9 260840 103667 -.72 .01|1.04 8.79| .99 -1.88|S .51 .44| 67.2 63.6| .00| ENGCLC9 |**

**| 18 211365 103577 .62A .00|1.00 1.14| .98 -3.74|T .51 .54| 56.2 54.4| -.08| ENGPAR8 |**

**| 7 224036 93587 -1.07A .01| .97 -6.42| .95 -7.36|U .45 .43| 61.8 58.9| .07| ENGCLC7 |**

**| 34 103357 50759 .64 .01| .96 -6.38| .93 -9.90|u .57 .52| 56.9 54.3| .00| ENGPAR25|**

**| 15 99258 56740 .52A .01| .87 -9.90| .92 -9.90|t .46 .50| 55.0 50.7| -.06| ENGPAR5 |**

**| 5 190020 93668 -.22A .00| .90 -9.90| .91 -9.90|s .53 .48| 61.0 55.2| .11| ENGCLC5 |**

**| 22 62103 37105 .64A .01| .87 -9.90| .88 -9.90|r .53 .52| 54.2 49.5| -.06| ENGPAR12|**

**| 25 216721 103508 .41 .00| .88 -9.90| .85 -9.90|q .59 .52| 61.4 55.4| .01| ENGPAR15|**

**| 26 98791 52722 .78 .01| .88 -9.90| .87 -9.90|p .59 .55| 57.3 52.7| .00| ENGPAR16|**

**| 14 332538 197274 .99A .00| .86 -9.90| .87 -9.90|o .60 .56| 54.9 49.6| -.07| ENGPAR4 |**

**| 27 95835 56410 .55 .01| .87 -9.90| .87 -9.90|n .57 .50| 54.5 50.6| .00| ENGPAR17|**

**| 41 217576 93732 -.94A .01| .86 -9.90| .83 -9.90|m .56 .44| 66.5 57.6| .14| ENGREL14|**

**| 39 203842 94162 -.47A .01| .80 -9.90| .83 -9.90|l .52 .47| 63.6 55.9| .05| ENGREL6 |**

**| 36 191077 103722 .94 .00| .76 -9.90| .79 -9.90|k .54 .55| 63.2 52.5| .00| ENGREL2 |**

**| 20 87018 56631 .97A .01| .75 -9.90| .78 -9.90|j .55 .51| 55.9 47.0| -.12| ENGPAR10|**

**| 28 153684 93506 .64 .00| .78 -9.90| .78 -9.90|i .59 .51| 56.6 49.7| .00| ENGPAR18|**

**| 42 216362 103516 .42 .00| .76 -9.90| .76 -9.90|h .53 .52| 64.5 55.3| .00| ENGREL15|**

**| 23 256367 160014 1.15 .00| .73 -9.90| .75 -9.90|g .57 .57| 59.6 49.2| .00| ENGPAR13|**

**| 35 152273 94117 .75A .00| .70 -9.90| .74 -9.90|f .50 .52| 59.1 48.9| -.06| ENGREL1 |**

**| 32 73180 37050 -.01 .01| .72 -9.90| .71 -9.90|e .58 .49| 64.7 54.5| .00| ENGPAR22|**

**| 38 196031 103516 .86A .00| .69 -9.90| .71 -9.90|d .56 .55| 65.0 52.8| -.02| ENGREL4 |**

**| 30 100713 56432 .39 .01| .69 -9.90| .69 -9.90|c .58 .50| 61.6 51.5| .00| ENGPAR20|**

**| 40 193156 93560 -.17A .00| .68 -9.90| .67 -9.90|b .61 .48| 65.7 55.1| -.02| ENGREL13|**

**| 29 112956 56410 -.04 .01| .63 -9.90| .63 -9.90|a .57 .48| 67.6 54.3| .00| ENGPAR19|**

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

**| MEAN 168435 83881 .16 .01| .97 -1.01| .97 -2.25| | 58.4 54.6| .00| |**

**| P.SD 72040.6 33174 .76 .00| .22 9.42| .22 8.63| | 7.4 5.4| .00| |**

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

**1Items color coded yellow are new deeper learning items added to survey in 2022; Point-To-Measures (PTM) color coded red demarcate the lower and upper bound of the PTMs.**

**Appendix C3, (Content validity): Technical quality (mean-square error) of 34-item Safety dimension scale1**

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

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

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

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

**| 49 99575 50775 .80 .01|1.95 9.90|1.84 9.90|A .53 .60| 35.5 54.1| .00| SAFBUL7 |**

**| 52 93503 37161 -1.55 .01|1.88 9.90|1.57 9.90|B .45 .44| 59.9 63.5| .00| SAFBUL10|**

**| 75 191861 103727 .91 .00|1.88 9.90|1.83 9.90|C .50 .62| 34.9 52.5| .00| SAFPSF7 |**

**| 53 93105 37135 -1.52 .01|1.86 9.90|1.62 9.90|D .41 .44| 58.6 63.3| .00| SAFBUL11|**

**| 47 120344 56638 -.51 .01|1.82 9.90|1.72 9.90|E .50 .53| 37.6 55.7| .00| SAFBUL5 |**

**| 44 101714 52747 .63 .01|1.81 9.90|1.73 9.90|F .55 .61| 37.1 53.4| .00| SAFBUL2 |**

**| 54 219221 103536 .33 .00|1.68 9.90|1.62 9.90|G .45 .58| 46.6 56.3| .00| SAFBUL12|**

**| 59 108398 50761 .41 .01|1.68 9.90|1.64 9.90|H .46 .57| 44.4 56.0| .00| SAFBUL17|**

**| 73 139482 56459 -1.43 .01|1.68 9.90|1.48 9.90|I .46 .47| 57.9 62.2| .00| SAFPSF4 |**

**| 68 110325 93915 1.48A .00|1.45 9.90|1.60 9.90|J .40 .59| 42.3 47.3| -.14| SAFEMO11|**

**| 50 252531 103720 -.52 .01|1.50 9.90|1.55 9.90|K .42 .52| 59.0 63.1| .00| SAFBUL8 |**

**| 63 88242 37340 -1.09 .01|1.55 9.90|1.47 9.90|L .40 .47| 53.5 59.4| .00| SAFEMO6 |**

**| 74 88163 37229 -1.10 .01|1.47 9.90|1.33 9.90|M .51 .47| 55.1 59.5| .00| SAFPSF5 |**

**| 62 145561 93536 .84A .00|1.41 9.90|1.41 9.90|N .49 .58| 41.5 48.0| -.19| SAFEMO4 |**

**| 48 135945 50825 -1.27 .01|1.37 9.90|1.19 9.90|O .45 .44| 75.8 73.5| .00| SAFBUL6 |**

**| 56 101616 56453 .17 .01|1.37 9.90|1.33 9.90|P .62 .56| 41.0 51.7| .00| SAFBUL14|**

**| 65 199746 103561 .70A .00|1.37 9.90|1.34 9.90|Q .54 .61| 47.6 54.0| .05| SAFEMO8 |**

**| 70 160898 103496 1.50 .00|1.29 9.90|1.27 9.90|q .59 .64| 45.6 49.8| .00| SAFEMO13|**

**| 71 256814 103766 -.86A .01|1.22 9.90|1.12 9.90|p .55 .50| 66.5 66.9| .22| SAFPSF1 |**

**| 55 356828 197649 .61A .00|1.16 9.90|1.17 9.90|o .56 .61| 51.1 52.1| -.02| SAFBUL13|**

**| 61 240112 103646 -.10A .01|1.09 9.90|1.00 .05|n .59 .55| 59.0 59.3| -.08| SAFEMO3 |**

**| 76 75927 37176 -.30 .01|1.08 9.90|1.04 5.42|m .56 .53| 55.1 54.8| .00| SAFPSF8 |**

**| 58 75273 37069 -.28 .01|1.07 9.12|1.04 5.00|l .56 .53| 55.5 54.8| .00| SAFBUL16|**

**| 51 445088 197255 -.54A .00|1.02 4.93| .95 -9.90|k .61 .54| 62.9 59.9| .12| SAFBUL9 |**

**| 60 432850 198072 -.04A .00| .96 -9.90| .99 -3.75|j .53 .58| 59.6 56.2| -.21| SAFEMO1 |**

**| 69 115852 50834 .05 .01| .94 -8.62| .98 -2.92|i .49 .55| 59.2 58.9| .00| SAFEMO12|**

**| 43 341225 146535 -.94A .00| .96 -9.11| .90 -9.90|h .59 .51| 65.7 61.4| .16| SAFBUL1 |**

**| 57 107712 56485 -.04 .01| .96 -6.86| .94 -9.90|g .57 .55| 56.9 53.2| .00| SAFBUL15|**

**| 67 191197 109332 .63A .00| .94 -9.90| .96 -8.11|f .54 .62| 54.2 50.9| -.02| SAFEMO10|**

**| 45 189451 94037 -.45A .00| .94 -9.90| .94 -9.90|e .57 .53| 59.3 55.4| .19| SAFBUL3 |**

**| 46 342971 197078 .63A .00| .93 -9.90| .94 -9.90|d .62 .61| 55.1 51.5| .09| SAFBUL4 |**

**| 72 93778 56516 .44 .01| .89 -9.90| .89 -9.90|c .58 .58| 55.9 49.9| .00| SAFPSF3 |**

**| 66 59221 37066 .77A .01| .85 -9.90| .85 -9.90|b .53 .58| 55.5 47.7| -.19| SAFEMO9 |**

**| 64 76326 37510 -.36A .01| .80 -9.90| .79 -9.90|a .58 .52| 63.6 55.1| .07| SAFEMO7 |**

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

**| MEAN 172084 84972 -.06 .01|1.32 4.06|1.27 3.37| | 53.2 56.2| .00| |**

**| P.SD 103776 50214 .83 .00| .35 8.74| .32 8.58| | 9.6 5.7| .00| |**

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

**1Items color coded yellow are new deeper learning items added to survey in 2022; Point-To-Measures (PTM) color coded red demarcate the lower and upper bound of the PTMs.**

**Appendix C4, (Content validity): Technical quality (mean-square error) of 37-item Environment dimension scale1**

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

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

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

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

**| 113 51861 37173 1.17 .01|1.48 9.90|1.80 9.90|A .40 .58| 40.5 47.1| .00| ENVMEN9 |**

**| 79 59076 50838 2.48 .01|1.45 9.90|1.55 9.90|B .47 .61| 42.0 48.2| .01| ENVDIS3 |**

**| 107 115243 52762 -.08A .01|1.34 9.90|1.28 9.90|C .53 .54| 55.7 58.1| .21| ENVMEN1 |**

**| 78 243990 103680 -.30A .01|1.29 9.90|1.25 9.90|D .54 .52| 60.9 61.0| .15| ENVDIS2 |**

**| 93 204292 103781 .77 .00|1.13 9.90|1.27 9.90|E .29 .58| 55.6 53.6| .00| ENVINS10|**

**| 97 147080 103545 1.84 .00|1.24 9.90|1.26 9.90|F .56 .61| 45.6 48.7| .00| ENVINS14|**

**| 90 202409 93752 -.64A .01|1.16 9.90|1.25 9.90|G .33 .50| 57.6 57.2| .22| ENVINS5 |**

**| 108 114384 50786 .28 .01|1.21 9.90|1.15 9.90|H .53 .54| 55.5 57.2| .00| ENVMEN2 |**

**| 77 175088 146555 1.90A .00|1.16 9.90|1.20 9.90|I .54 .61| 48.5 48.1| -.20| ENVDIS1 |**

**| 85 221700 103499 .38 .00|1.12 9.90|1.09 9.90|J .59 .56| 56.5 55.9| .00| ENVDIS9 |**

**| 101 113827 52735 .19 .01|1.12 9.90|1.09 9.90|K .55 .55| 54.2 55.8| .00| ENVINS18|**

**| 111 75454 37310 -.39A .01|1.10 9.90|1.09 9.74|L .59 .51| 59.5 56.3| .35| ENVMEN6 |**

**| 81 119297 50883 .03 .01|1.08 9.90|1.05 6.78|M .51 .52| 59.2 58.9| .00| ENVDIS5 |**

**| 88 265632 103520 -1.15A .01|1.08 9.90| .93 -9.09|N .60 .45| 72.8 71.5| .33| ENVINS3 |**

**| 106 126562 50759 -.41 .01|1.08 9.90|1.00 .54|O .52 .49| 65.9 64.1| .00| ENVINS23|**

**| 100 234908 103489 .07 .01|1.06 9.90|1.03 5.76|P .55 .54| 59.1 57.8| .00| ENVINS17|**

**| 110 107970 56594 -.03A .01|1.05 9.12|1.03 5.34|Q .60 .53| 53.7 53.5| .12| ENVMEN4 |**

**| 92 107258 56767 .44A .01| .93 -9.90|1.03 4.51|R .31 .55| 56.4 50.9| -.30| ENVINS9 |**

**| 109 110622 56418 -.16A .01|1.01 1.84|1.00 -.46|S .58 .53| 56.3 54.4| .14| ENVMEN3 |**

**| 91 137522 56726 -1.36A .01| .98 -2.54| .96 -5.09|r .56 .46| 66.9 62.0| .21| ENVINS8 |**

**| 80 113977 56412 -.31A .01| .96 -6.88| .95 -8.36|q .54 .52| 60.3 55.5| .16| ENVDIS4 |**

**| 82 69042 37333 .43A .01| .96 -6.12| .96 -4.74|p .59 .56| 52.8 51.8| -.11| ENVDIS6 |**

**| 112 229363 103592 .24A .00| .94 -9.90| .91 -9.90|o .60 .55| 60.1 56.6| -.03| ENVMEN7 |**

**| 87 406449 160345 -1.03A .00| .93 -9.90| .82 -9.90|n .60 .49| 72.9 66.1| .00| ENVINS2 |**

**| 89 110772 50841 .46 .01| .87 -9.90| .86 -9.90|m .53 .55| 62.2 55.8| .00| ENVINS4 |**

**| 86 288857 146742 .22A .00| .83 -9.90| .86 -9.90|l .49 .56| 60.6 53.8| -.01| ENVINS1 |**

**| 104 70258 37070 .23 .01| .85 -9.90| .85 -9.90|k .62 .55| 58.9 53.4| .00| ENVINS21|**

**| 83 268708 146966 .63A .00| .81 -9.90| .81 -9.90|j .65 .58| 57.3 51.3| -.12| ENVDIS7 |**

**| 99 94013 52749 1.00 .01| .79 -9.90| .80 -9.90|i .61 .59| 60.6 51.9| .00| ENVINS16|**

**| 98 75920 37492 -.25A .01| .78 -9.90| .79 -9.90|h .62 .52| 64.6 55.6| .20| ENVINS15|**

**| 95 146900 93595 .79A .00| .75 -9.90| .78 -9.90|g .62 .57| 57.9 48.8| -.01| ENVINS12|**

**| 96 63679 37155 .76A .01| .78 -9.90| .78 -9.90|f .63 .57| 58.4 50.1| -.17| ENVINS13|**

**| 84 155411 93493 .61 .00| .76 -9.90| .76 -9.90|e .64 .56| 58.2 50.0| .00| ENVDIS8 |**

**| 102 62286 37067 .65 .01| .74 -9.90| .76 -9.90|d .64 .57| 59.8 50.2| .00| ENVINS19|**

**| 94 445325 197206 -.13A .00| .74 -9.90| .73 -9.90|c .66 .55| 65.6 57.2| -.15| ENVINS11|**

**| 105 116505 56446 -.24 .01| .69 -9.90| .70 -9.90|b .63 .52| 66.1 55.2| .00| ENVINS22|**

**| 103 61049 37028 .71 .01| .66 -9.90| .67 -9.90|a .66 .57| 61.8 50.1| .00| ENVINS20|**

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

**| MEAN 154397 77111 .27 .01|1.00 -.12|1.00 -.94| | 58.4 55.0| .00| |**

**| P.SD 92334.4 40980 .78 .00| .21 9.46| .24 8.91| | 6.8 5.2| .00| |**

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

**1Items color coded yellow are new deeper learning items added to survey in 2022; Point-To-Measures (PTM) color coded red demarcate the lower and upper bound of the PTMs.**

**Appendix C5, (Content validity): Technical quality (mnsq) of 24-item Engagement Participation topic scale1**

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

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

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

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

**| 11 236056 103667 -.04 .01|1.65 9.90|1.63 9.90|A .41 .52| 50.5 58.9| .00| ENGPAR1 |**

**| 13 85845 37412 -.74 .01|1.27 9.90|1.28 9.90|B .42 .47| 52.3 57.0| .00| ENGPAR3 |**

**| 19 244602 103761 -.37A .01|1.24 9.90|1.27 9.90|C .42 .50| 57.5 61.1| .11| ENGPAR9 |**

**| 12 124418 56162 -.59A .01|1.19 9.90|1.17 9.90|D .46 .47| 56.9 56.5| .13| ENGPAR2 |**

**| 21 67911 37467 .15A .01|1.06 7.95|1.09 9.90|E .54 .52| 51.8 54.1| .17| ENGPAR11|**

**| 24 101114 93496 1.70 .00|1.01 1.73|1.00 .85|F .59 .55| 45.7 45.9| .00| ENGPAR14|**

**| 16 251838 103652 -.54A .01|1.00 -.28| .93 -9.90|G .55 .48| 65.3 62.3| .07| ENGPAR6 |**

**| 31 159829 93295 .56 .00| .89 -9.90| .96 -8.02|H .41 .53| 56.9 50.9| .00| ENGPAR21|**

**| 17 207405 103532 .61A .00| .95 -9.90| .90 -9.90|I .65 .57| 57.3 54.9| -.01| ENGPAR7 |**

**| 33 48825 37030 1.24 .01| .89 -9.90| .90 -9.90|J .59 .56| 51.7 46.8| .00| ENGPAR24|**

**| 34 103357 50759 .61 .01| .90 -9.90| .87 -9.90|K .61 .56| 60.2 55.5| .00| ENGPAR25|**

**| 18 211365 103577 .62A .00| .89 -9.90| .86 -9.90|L .58 .57| 59.7 54.9| -.10| ENGPAR8 |**

**| 15 99258 56740 .52A .01| .83 -9.90| .88 -9.90|l .51 .52| 57.2 51.4| .00| ENGPAR5 |**

**| 23 256367 160014 1.18 .00| .81 -9.90| .85 -9.90|k .54 .59| 57.5 49.5| .00| ENGPAR13|**

**| 25 216721 103508 .40 .00| .85 -9.90| .82 -9.90|j .61 .56| 62.1 55.7| .00| ENGPAR15|**

**| 26 98791 52722 .80 .01| .82 -9.90| .81 -9.90|i .63 .58| 59.4 53.0| .00| ENGPAR16|**

**| 22 62103 37105 .64A .01| .80 -9.90| .81 -9.90|h .58 .54| 57.3 50.4| -.04| ENGPAR12|**

**| 14 332538 197274 .99A .00| .78 -9.90| .79 -9.90|g .65 .59| 58.5 50.2| -.06| ENGPAR4 |**

**| 27 95835 56410 .62 .01| .78 -9.90| .78 -9.90|f .63 .52| 57.0 50.2| .00| ENGPAR17|**

**| 20 87018 56631 .97A .01| .68 -9.90| .70 -9.90|e .61 .53| 60.0 47.9| -.06| ENGPAR10|**

**| 28 153684 93506 .69 .00| .70 -9.90| .70 -9.90|d .64 .53| 59.8 50.2| .00| ENGPAR18|**

**| 32 73180 37050 .00 .01| .70 -9.90| .69 -9.90|c .60 .51| 65.2 54.2| .00| ENGPAR22|**

**| 30 100713 56432 .46 .01| .62 -9.90| .62 -9.90|b .63 .51| 64.4 51.5| .00| ENGPAR20|**

**| 29 112956 56410 .02 .01| .61 -9.90| .61 -9.90|a .59 .50| 68.7 54.3| .00| ENGPAR19|**

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

**| MEAN 147155 78651 .44 .01| .91 -4.97| .91 -5.25| | 58.0 53.2| .00| |**

**| P.SD 76000.5 39831 .59 .00| .23 7.99| .23 8.06| | 5.1 4.1| .00| |**

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

**1Items color coded yellow are new deeper learning items added to survey in 2022; Point-To-Measures (PTM) color coded red demarcate the lower and upper bound of the PTMs.**

**Appendix C6, (Content validity): Technical quality (mnsq) of 17-item Safety Bullying topic scale**

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

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

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

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

**| 49 99575 50775 .71 .01|1.78 9.90|1.61 9.90|A .60 .65| 39.5 54.7| .00| SAFBUL7 |**

**| 52 93503 37161 -1.68 .01|1.73 9.90|1.36 9.90|B .50 .49| 64.8 65.9| .00| SAFBUL10|**

**| 53 93105 37135 -1.66 .01|1.72 9.90|1.42 9.90|C .47 .49| 59.4 64.5| .00| SAFBUL11|**

**| 47 120344 56638 -.65 .01|1.67 9.90|1.54 9.90|D .56 .59| 42.0 56.9| .00| SAFBUL5 |**

**| 59 108398 50761 .31 .01|1.67 9.90|1.57 9.90|E .50 .62| 47.0 58.0| .00| SAFBUL17|**

**| 44 101714 52747 .60 .01|1.59 9.90|1.44 9.90|F .63 .66| 42.8 55.3| .00| SAFBUL2 |**

**| 54 219221 103536 .25 .00|1.47 9.90|1.34 9.90|G .55 .63| 51.7 58.1| .00| SAFBUL12|**

**| 50 252531 103720 -.62 .01|1.45 9.90|1.39 9.90|H .47 .56| 61.2 65.4| .00| SAFBUL8 |**

**| 48 135945 50825 -1.42 .01|1.33 9.90|1.09 5.36|I .47 .47| 76.2 74.6| .00| SAFBUL6 |**

**| 56 101616 56453 .06 .01|1.22 9.90|1.16 9.90|h .67 .62| 46.6 52.1| .00| SAFBUL14|**

**| 55 356828 197649 .61A .00|1.03 8.49|1.00 .46|g .64 .67| 54.6 52.8| -.10| SAFBUL13|**

**| 46 342971 197078 .63A .00| .97 -8.53| .97 -7.49|f .63 .67| 55.8 52.8| .01| SAFBUL4 |**

**| 51 445088 197255 -.54A .00| .97 -9.21| .89 -9.90|e .62 .60| 63.1 60.5| .00| SAFBUL9 |**

**| 57 107712 56485 -.16 .01| .92 -9.90| .89 -9.90|d .61 .61| 59.1 54.4| .00| SAFBUL15|**

**| 58 75273 37069 -.38 .01| .91 -9.90| .87 -9.90|c .64 .58| 59.6 54.9| .00| SAFBUL16|**

**| 43 341225 146535 -.94A .00| .90 -9.90| .84 -9.90|b .60 .57| 66.7 61.9| .04| SAFBUL1 |**

**| 45 189451 94037 -.45A .00| .88 -9.90| .86 -9.90|a .60 .59| 60.9 55.5| .07| SAFBUL3 |**

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

**| MEAN 187324 89756 -.31 .01|1.31 2.95|1.19 2.23| | 55.9 58.7| .00| |**

**| P.SD 113930 57339 .77 .00| .34 9.25| .27 8.99| | 9.6 5.9| .00| |**

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

**Appendix C7, (Content validity): Technical quality (mnsq) of 21-item Environment Instructional topic scale**

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

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

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

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

**| 93 204292 103781 .76 .00|1.06 9.90|1.15 9.90|A .36 .59| 57.2 53.7| .00| ENVINS10|**

**| 97 147080 103545 1.84 .00|1.13 9.90|1.15 9.90|B .61 .63| 48.8 49.5| .00| ENVINS14|**

**| 90 202409 93752 -.64A .01|1.03 6.91|1.08 9.90|C .42 .50| 60.8 57.6| .22| ENVINS5 |**

**| 101 113827 52735 .17 .01|1.02 3.54| .99 -2.06|D .60 .56| 58.4 56.6| .00| ENVINS18|**

**| 88 265632 103520 -1.15A .01|1.01 2.20| .86 -9.90|E .62 .46| 73.5 71.7| .31| ENVINS3 |**

**| 106 126562 50759 -.42 .01| .99 -1.66| .89 -9.90|F .57 .51| 68.2 64.2| .01| ENVINS23|**

**| 100 234908 103489 .05 .01| .98 -3.65| .94 -9.90|G .59 .55| 61.9 58.7| .00| ENVINS17|**

**| 91 137522 56726 -1.36A .01| .94 -9.90| .88 -9.90|H .59 .45| 70.5 61.6| .23| ENVINS8 |**

**| 87 406449 160345 -1.03A .00| .88 -9.90| .76 -9.90|I .62 .49| 74.2 66.0| .00| ENVINS2 |**

**| 92 107258 56767 .44A .01| .83 -9.90| .88 -9.90|J .39 .55| 60.0 51.2| -.29| ENVINS9 |**

**| 89 110772 50841 .46 .01| .85 -9.90| .84 -9.90|K .56 .57| 63.9 57.2| .00| ENVINS4 |**

**| 86 288857 146742 .22A .00| .78 -9.90| .81 -9.90|j .53 .57| 63.2 54.0| -.01| ENVINS1 |**

**| 104 70258 37070 .20 .01| .79 -9.90| .78 -9.90|i .64 .56| 60.3 53.5| .00| ENVINS21|**

**| 98 75920 37492 -.25A .01| .74 -9.90| .77 -9.90|h .63 .53| 65.8 56.5| .17| ENVINS15|**

**| 99 94013 52749 .98 .01| .76 -9.90| .76 -9.90|g .63 .60| 60.3 51.0| .00| ENVINS16|**

**| 96 63679 37155 .76A .01| .73 -9.90| .73 -9.90|f .65 .58| 58.3 49.4| -.20| ENVINS13|**

**| 95 146900 93595 .79A .00| .71 -9.90| .72 -9.90|e .64 .57| 59.0 48.7| -.01| ENVINS12|**

**| 102 62286 37067 .62 .01| .71 -9.90| .72 -9.90|d .65 .58| 61.6 50.8| .00| ENVINS19|**

**| 94 445325 197206 -.13A .00| .71 -9.90| .69 -9.90|c .67 .56| 64.9 57.0| -.16| ENVINS11|**

**| 105 116505 56446 -.22 .01| .66 -9.90| .65 -9.90|b .65 .52| 67.5 55.7| .00| ENVINS22|**

**| 103 61049 37028 .68 .01| .62 -9.90| .62 -9.90|a .68 .58| 64.4 50.7| .00| ENVINS20|**

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

**| MEAN 165786 79467 .13 .01| .85 -5.31| .84 -6.70| | 63.0 56.0| .00| |**

**| P.SD 106428 44316 .76 .00| .15 7.10| .15 6.98| | 5.7 5.8| .00| |**

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

**1Items color coded yellow are new deeper learning items added to survey in 2022; Point-To-Measures (PTM) color coded red demarcate the lower and upper bound of the PTMs.**

**Appendix D, (Structural Validity): Winsteps Residual Analyses**

**D1: Residual analyses of 113-item VOCAL data (Grades 4, 5, 8, and 10 combined)**

|  |  |  |
| --- | --- | --- |
| **Variance Component** | **Eigenvalue** | **Observed (%)** |
| Raw variance explained by measures | 64.8 | 36.5% |
| Raw variance explained by persons | 38.0 | 21.4% |
| Raw Variance explained by items | 26.8 | 15.1% |
| Unexplained variance in 1st contrast | 4.2 | 2.4% |
| Item variance to 1st contrast multiple |  | 6.3x |

**D2, (Structural Validity): Residual analyses of dimension data (Grades 4, 5, 8, and 10 combined)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variance Component** | **Engagement**  **42 items** | **Safety**  **34 items1** | **Environment**  **37 items** |
| Raw variance explained by measures | 41.7% | 38.7% | 44.3% |
| Raw variance explained by persons | 25.0% | 26.2% | 26.5% |
| Raw Variance explained by items | 16.7% | 12.5% | 17.8% |
| Unexplained variance in 1st contrast | 3.2% | 4.7% | 2.5% |
| Eigenvalue 1st contrast | 2.3 | 2.6 | 1.7 |
| 1st contrast multiple | 5.2x | 2.7x | 7.1x |

1Bullying protective factor items (BUL1, BUL3, BUL4, and BUL9) separated from the bullying behavior items (BUL2, BUL5, , BUL7, BUL10 to BUL16)

**D3, (Structural Validity): Residual analyses of topic data (Grades 4, 5, 8, and 10 combined)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variance Component** | **Engagement**  **Participation**  **24 items** | **Safety**  **Bullying**  **17 items1** | **Environment**  **Instructional**  **21 items** |
| Raw variance explained by measures | 43.8% | 43.2% | 46.5% |
| Raw variance explained by persons | 32.1% | 33.1% | 29.5% |
| Raw Variance explained by items | 11.8% | 10.1% | 17.0% |
| Unexplained variance in 1st contrast | 3.3% | 7.1% | 4.2% |
| Eigenvalue 1st contrast | 1.4 | 2.1 | 1.6 |
| 1st contrast multiple | 3.6x | 1.4x | 4.0x |

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

**Appendix E, (Structural Validity): Pearson Correlations between student dimension/topic scores**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Indicator1** | **scl22ms2** | **eng22ms3** | **saf22ms4** | **env22ms5** | **par22ms6** | **bul22ms7** | **ins22ms8** |
| **scl22ms** | R | 1 |
| Sig. |  |
| N | 198169 |
| **eng22ms** | R | .928\*\* | 1 |
| Sig. | 0.000 |  |
| N | 198045 | 198045 |
| **saf22ms** | R | .880\*\* | .695\*\* | 1 |
| Sig. | 0.000 | 0.000 |  |
| N | 198157 | 198040 | 198157 |
| **env22ms** | R | .927\*\* | .846\*\* | .710\*\* | 1 |
| Sig. | 0.000 | 0.000 | 0.000 |  |
| N | 198106 | 198039 | 198099 | 198106 |
| **par22ms** | R | .828\*\* | .932\*\* | .561\*\* | .777\*\* | 1 |
| Sig. | 0.000 | 0.000 | 0.000 | 0.000 |  |
| N | 198036 | 198036 | 198034 | 198033 | 198036 |
| **bul22ms** | R | .771\*\* | .594\*\* | .917\*\* | .598\*\* | .460\*\* | 1 |
| Sig. | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |  |
| N | 197887 | 197884 | 197887 | 197884 | 197880 | 197887 |
| **ins22ms** | R | .865\*\* | .809\*\* | .643\*\* | .934\*\* | .755\*\* | .533\*\* | 1 |
| Sig. | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |  |
| N | 198102 | 198036 | 198096 | 198102 | 198031 | 197883 | 198102 |

1R=Pearson Correlation, Sig.= \*\**p*<0.01, N=Number of students; 2Overall VOCAL score; 3Engagement dimension score; 4Safety dimension score; 5Environment dimension score; 6Participation topic score; 7Bullying topic score; 8Instructional topic score

**Appendix F, (Substantive Validity): Rating scale structure for 113-item VOCAL instrument (all 4 grades jointly calibrated)**

**SUMMARY OF CATEGORY STRUCTURE. Model="R"**

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

**|CATEGORY OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|**

**|LABEL SCORE COUNT %|AVRGE EXPECT| MNSQ MNSQ||THRESHOLD| MEASURE|**

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

**| 0 0 660636 7| -.42 -.71| 1.30 1.57|| NONE A |( -2.62)| 0**

**| 1 1 1624723 18| .09 .08| 1.00 1.01|| -1.28A| -.92 | 1**

**| 2 2 3926310 42| .86 .91| .97 .88|| -.43A| .76 | 2**

**| 3 3 3053494 33| 1.91 1.91| 1.07 1.07|| 1.71A|( 2.89)| 3**

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

**| MISSING 13128E3 59| 1.05 | || | |**

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

**OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.**

**CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections**

**P -+---------+---------+---------+---------+---------+---------+-**

**R 1.0 + +**

**O | |**

**B | |**

**A |0 |**

**B .8 + 000 3+**

**I | 000 33 |**

**L | 00 333 |**

**I | 00 33 |**

**T .6 + 0 22 33 +**

**Y | 00 22222 22222 3 |**

**.5 + 00 222 222 33 +**

**O | 0 22 3\*22 |**

**F .4 + 11\*\*1111111\* 33 22 +**

**| 111 0 22 111 33 222 |**

**R | 111 0022 111 33 22 |**

**E | 111 2200 11 33 222 |**

**S .2 + 1111 22 00 1\*\* 2+**

**P |11 222 00 333 111 |**

**O | 222 0\*\*\*3 1111 |**

**N | 2222222 333333 000000 11111111 |**

**S .0 +\*\*\*\*\*333333333333333333 000000000000000\*\*\*\*\*\*\*\*+**

**E -+---------+---------+---------+---------+---------+---------+-**

**-3 -2 -1 0 1 2 3**

**PERSON [MINUS] ITEM MEASURE**

**Appendix G1, (Substantive Validity): Item-variable map for 2022 VOCAL survey items with dimensions color coded1**

This graphic shows the item-variable map for the 113-item VOCAL school climate scale. Items color coded yellow are from the engagement dimension; items in pink belong to the safety dimension; items in green belong to the environment dimension.

1Engagement dimension items are in yellow; Safety dimension items are in pink; Environment dimension items are in green; Items with purple text indicate new deeper learning items added in 2022

**Appendix G2, (Substantive Validity): Item-variable map for 2022 VOCAL survey items with topics areas color coded1**

**This graphic shows the item-variable map for the 113-item VOCAL school climate scale. Items with the first three letters color coded yellow are from the engagement dimension; items with the first three letters in pink belong to the safety dimension; items with the first three letters in green belong to the environment dimension.**

1Engagement dimension items (ENG, yellow) are further color coded by topic (Culture Competence, CLC; Participation, PAR; Relationships, REL); Safety dimension items (SAF, pink) are further color coded by topic (Emotional safety, EMO; Physical safety, PSF; Bullying, BUL); Environment dimension items (ENV, green) are further color coded by topic (Instructional environment, INS; Mental-health environment, MEN; Discipline environment, DIS

**Appendix G3, (Substantive Validity): Item-threshold map for 2022 VOCAL survey items**

**MEASURE | BOTTOM P=50% | MEASURE | TOP P=50% MEASURE**

**<more> ----- PERSON -+- ITEM -+- ITEM -+- ITEM <rare>**

**8 . + + + 8**

**| | |**

**| | |**

**. | | |**

**7 + + + 7**

**| | |**

**| | |**

**| | |**

**6 . + + + 6**

**. | | |**

**| | |**

**. | | |**

**5 . + + + 5**

**. | | |**

**. | | |**

**. | | |**

**4 . + + + . 4**

**. | | | .**

**. | | | .**

**.# | | | ##**

**3 .## + + + . 3**

**.## | | | ###.**

**.#### | | | ########**

**.###### | | . | ########.**

**2 .####### + + . + #####. 2**

**.######### | | . | #####.**

**.########## | | ## | ######.**

**.########### | | . | #####**

**1 .############ + + ###. + ## 1**

**.############ | . | ######## | ####**

**.########### | . | ########. | ###.**

**.######### | . | #####. | .**

**0 .###### + ## + #####. + 0**

**.#### | . | ######. |**

**.## | ###. | ##### |**

**.# | ######## | ## |**

**-1 . + ########. + #### + -1**

**. | #####. | ###. |**

**. | #####. | . |**

**. | ######. | |**

**-2 . + ##### + + -2**

**. | ## | |**

**. | #### | |**

**. | ###. | |**

**-3 . + . + + -3**

**. | | |**

**. | | |**

**. | | |**

**-4 . + + + -4**

**. | | |**

**| | |**

**. | | |**

**-5 + + + -5**

**| | |**

**| | |**

**| | |**

**-6 + + + -6**

**. | | |**

**. | | |**

**| | |**

**-7 . + + + -7**

**<less> ----- PERSON -+- ITEM -+- ITEM -+- ITEM <freq>**

**WAAAA EACH "#" IN THE PERSON COLUMN IS 1622 PERSON: EACH "." IS 1 TO 1621**

**XAAAA EACH "#" IN THE ITEM COLUMN IS 2 ITEM: EACH "." IS 1**

**Appendix H1, (Generalizability): Person Reliability of overall VOCAL scale and by grade-level**

**School Climate (all persons = 198,169; items = 113)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR)2 | Person Separation Index (PSI: G) | Person Strata (H) | Mean ±SD**2** |
| Real – Model**1** | 0.93 – 0.94 | 3.58 – 4.10 | 5.1 – 5.8 | 1.14 ± 0.99 |

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

**Grade 4 items (persons = 50,933; items = 46)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.90 – 0.93 | 3.07 – 3.63 | 4.4 – 5.2 | 1.65± 0.95 |

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

**Grade 5 items (persons =52,878; items = 46)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.92 – 0.94 | 3.47 – 3.96 | 5.0 – 5.6 | 1.38 ± 0.99 |

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

**Grade 8 items (persons =56,795; items = 48)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.92 – 0.93 | 3.29 – 3.68 | 4.7 – 5.2 | 0.72 ± 0.83 |

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

**Grade 10 items (persons =37,562; items = 48)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.92 – 0.93 | 3.31 – 3.76 | 4.7 – 5.3 | 0.78 ± 0.89 |

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

**Appendix H2, (Generalizability): Person Reliability of VOCAL dimension scales (all grades)**

**Engagement Dimension (all persons =198,045; items = 42)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata (H) | Mean ±SD**2** |
| Real - Model**1** | 0.84 – 0.86 | 2.25 – 2.48 | 3.3 – 3.6 | 1.19 ± 1.08 |

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

**Safety Dimension items (all persons = 198,157; items = 34)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.83– 0.86 | 2.17 – 2.53 | 3.2 – 3.7 | 1.13 ± 1.25 |

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

**Environment Dimension items (all persons = 198,106; items = 37)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.82 – 0.85 | 2.16 – 2.42 | 3.2 – 3.6 | 1.25 ± 1.17 |

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

**Appendix I3, (Generalizability): Person Reliability of VOCAL topic scales (all grades)**

**Engagement: Participation topic items (all persons = 198,036; items = 24)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.74 – 0.78 | 1.71 – 1.89 | 2.6 – 2.8 | 1.22 ± 1.16 |

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

**Safety: Bullying/Cyberbullying topic items (all persons = 197,887; items = 17)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.73– 0.78 | 1.63 – 1.86 | 2.5 – 2.8 | 1.11 ± 1.46 |

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

**Environment: Instructional Environment topic items (all persons = 198,080; items = 21)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Person Separation Reliability (PSR) | Person Separation Index (PSI: G) | Person Strata  (H) | Mean ±SD**2** |
| Real - Model**1** | 0.71 – 0.75 | 1.57 – 1.73 | 2.4 – 2.6 | 1.26 ± 1.22 |

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

**Appendix I1, (Generalizability): DIF by Gender Status1,2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item3** | **Female measure** | **Male measure** | **VOCAL**  **DIF Size** | **Dimension**  **DIF Size** |
| SAFBUL7 | 0.46 | 1.14 | -0.68 | -0.68 |
| SAFBUL2 | 0.36 | 0.93 | -0.58 |  |
| ENGCLC4 | -0.48 | 0.09 | -0.57 |  |
| SAFBUL11 | -1.57 | -1.18 | -0.39 |  |
| **ENGCLC10** | -1.18 | -0.8 | -0.38 |  |
| ENVINS5 | -0.59 | -0.32 | -0.27 |  |
| **ENGPAR22** | -0.16 | 0.09 | -0.25 |  |
| ENGCLC3 | 0.32 | 0.56 | -0.24 |  |
| **ENGPAR17** | 0.37 | 0.58 | -0.21 |  |
| ENVINS12 | 0.62 | 0.82 | -0.2 |  |
| SAFEMO3 | -0.22 | -0.03 | -0.19 |  |
| ENGCLC2 | -1.31 | -1.13 | -0.17 |  |
| ENVINS9 | 0.03 | 0.19 | -0.17 |  |
| ENVINS14 | 1.65 | 1.82 | -0.17 |  |
| **ENGPAR18** | 0.51 | 0.67 | -0.16 |  |
| ENVINS1 | 0.09 | 0.25 | -0.16 |  |
| ENGCLC6 | -0.53 | -0.38 | -0.15 |  |
| SAFEMO6 | -1.02 | -0.87 | -0.15 |  |
| **ENGPAR20** | 0.25 | 0.39 | -0.14 |  |
| ENVINS2 | -1.13 | -0.99 | -0.14 |  |
| ENVMEN3 | -0.1 | 0.04 | -0.14 |  |
| **ENGPAR19** | -0.18 | -0.05 | -0.13 |  |
| ENGREL3 | -1.45 | -1.32 | -0.13 |  |
| ENVDIS2 | -0.3 | -0.17 | -0.13 |  |
| **BETTER CONTRASTING ITEMS ARE OMITTED** | | | | |
| ENGCLC1 | -0.91 | -1.03 | 0.12 |  |
| ENVDIS6 | 0.28 | 0.16 | 0.12 |  |
| SAFPSF1 | -0.51 | -0.64 | 0.13 |  |
| ENGREL6 | -0.39 | -0.53 | 0.14 |  |
| ENGPAR14 | 1.65 | 1.5 | 0.15 |  |
| **ENGREL15** | 0.5 | 0.35 | 0.15 |  |
| SAFEMO11 | 1.48 | 1.33 | 0.15 |  |
| SAFBUL13 | 0.74 | 0.57 | 0.17 |  |
| ENGREL2 | 1.05 | 0.85 | 0.19 |  |
| SAFEMO7 | -0.06 | -0.26 | 0.2 |  |
| ENGCLC5 | -0.05 | -0.27 | 0.22 |  |
| ENGPAR13 | 1.24 | 1.01 | 0.23 |  |
| ENGREL1 | 0.75 | 0.51 | 0.24 |  |
| SAFBUL5 | -0.26 | -0.5 | 0.24 |  |
| **SAFEMO13** | 1.62 | 1.37 | 0.25 |  |
| SAFPSF7 | 1.06 | 0.81 | 0.25 |  |
| ENGREL4 | 0.97 | 0.72 | 0.26 |  |
| SAFPSF8 | -0.02 | -0.34 | 0.32 |  |
| ENVMEN9 | 1.23 | 0.87 | 0.37 |  |
| SAFPSF5 | -0.76 | -1.18 | 0.42 |  |

1This DIF analyses examines if the item has the same difficulty for the two comparison groups;2There is an insufficient number of non-binary students to include in the DIF analyses;3New DL items bolded

**Appendix I2, (Generalizability): DIF by Race/Ethnicity Status1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Item2** | **Group** | **COUNT** | **Without**  **DIF Measure** | **Group Measure** | **VOCAL**  **DIF size** | **Dimension**  **DIF Size** | |
| SAFBUL11 | White | 22305 | -1.37 | -2.01 | -0.64 | -0.59 | |
| ENGPAR8 | Afr./Black | 9094 | 0.62 | 0.07 | -0.55 |  |
| ENVINS13 | Afr./Black | 3328 | 0.76 | 0.24 | -0.52 |  |
| **ENVINS18** | Afr./Black | 4673 | 0.15 | -0.35 | -0.51 |  |
| ENVINS13 | Hispanic | 7946 | 0.76 | 0.28 | -0.48 |  |
| ENGPAR10 | Hispanic | 12894 | 0.97 | 0.5 | -0.47 |  |
| ENVDIS1 | Hispanic | 33341 | 1.9 | 1.43 | -0.47 |  |
| ENGCLC4 | Asian | 2119 | -0.19 | -0.64 | -0.46 |  |
| ENGPAR10 | Afr./Black | 5096 | 0.97 | 0.52 | -0.45 |  |
| SAFPSF4 | Asian | 3898 | -1.28 | -1.68 | -0.41 |  |
| SAFBUL5 | Asian | 3902 | -0.38 | -0.78 | -0.4 |  |
| ENGPAR8 | Hispanic | 24609 | 0.62 | 0.23 | -0.39 |  |
| ENGPAR14 | Hispanic | 20728 | 1.57 | 1.18 | -0.39 |  |
| **ENVINS18** | Hispanic | 12498 | 0.15 | -0.24 | -0.39 |  |
| **ENVINS23** | Afr./Black | 4410 | -0.52 | -0.9 | -0.39 |  |
| ENVDIS1 | Asian | 9732 | 1.9 | 1.52 | -0.38 |  |
| ENGPAR12 | Afr./Black | 3319 | 0.64 | 0.27 | -0.37 |  |
| ENVDIS7 | Asian | 9750 | 0.63 | 0.26 | -0.37 |  |
| ENVINS9 | Afr./Black | 5119 | 0.44 | 0.07 | -0.37 |  |
| ENGPAR4 | Afr./Black | 17507 | 0.99 | 0.63 | -0.36 |  |
| ENVDIS1 | Afr./Black | 13097 | 1.9 | 1.54 | -0.36 |  |
| ENVINS13 | Asian | 2112 | 0.76 | 0.4 | -0.36 |  |
| ENVINS9 | White | 32486 | 0.44 | 0.09 | -0.35 |  |
| ENGPAR14 | Afr./Black | 8383 | 1.57 | 1.23 | -0.34 |  |
| ENGPAR7 | Hispanic | 24591 | 0.61 | 0.28 | -0.33 |  |
| ENVDIS6 | Hispanic | 7994 | 0.43 | 0.1 | -0.33 |  |
| ENVINS13 | Other | 1441 | 0.76 | 0.43 | -0.33 |  |
| ENGPAR7 | Afr./Black | 9085 | 0.61 | 0.29 | -0.32 |  |
| ENGPAR12 | Hispanic | 7939 | 0.64 | 0.32 | -0.32 |  |
| ENVINS14 | Asian | 7433 | 1.73 | 1.41 | -0.32 |  |
| ENGPAR5 | Afr./Black | 5118 | 0.52 | 0.21 | -0.31 |  |
| ENVDIS1 | Other | 6228 | 1.9 | 1.59 | -0.31 |  |
| ENVDIS6 | Asian | 2120 | 0.43 | 0.12 | -0.31 |  |
| ENVINS9 | Hispanic | 12942 | 0.44 | 0.13 | -0.31 |  |
| **ENVINS21** | Afr./Black | 3316 | 0.13 | -0.18 | -0.31 |  |
| SAFBUL10 | Asian | 2108 | -1.4 | -1.7 | -0.3 |  |
| ENVDIS7 | Hispanic | 33472 | 0.63 | 0.33 | -0.3 |  |
| ENVINS11 | Afr./Black | 17492 | -0.13 | -0.43 | -0.3 |  |
| **ENVINS21** | Hispanic | 7928 | 0.13 | -0.17 | -0.3 |  |
| **BETTER CONTRASTING ITEMS ARE OMITTED** | | | | | | |
| SAFBUL1 | Hispanic | 33358 | -0.94 | -0.64 | 0.3 |  |
| SAFBUL2 | Afr./Black | 4675 | 0.66 | 0.96 | 0.3 |  |
| SAFBUL9 | Afr./Black | 17489 | -0.54 | -0.24 | 0.3 |  |
| SAFBUL13 | Afr./Black | 17553 | 0.61 | 0.91 | 0.3 |  |

**Appendix I2, (Generalizability): DIF by Race/Ethnicity Status continued1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Item2** | | **Group** | **COUNT** | **Without**  **DIF Measure** | **Group Measure** | **VOCALDIF size** | **Dimension DIF Size** |
| **ENVINS18** | Asian | | 3715 | 0.15 | 0.46 | 0.30 |  |
| ENVMEN6 | Hispanic | | 7996 | -0.39 | -0.09 | 0.30 |  |
| ENGPAR9 | Asian | | 7442 | -0.37 | -0.06 | 0.31 |  |
| SAFPSF3 | Afr./Black | | 5081 | 0.55 | 0.86 | 0.31 |  |
| ENGPAR1 | Hispanic | | 24636 | -0.02 | 0.3 | 0.32 |  |
| SAFEMO6 | Hispanic | | 7998 | -0.95 | -0.62 | 0.32 |  |
| SAFPSF4 | Afr./Black | | 5067 | -1.28 | -0.95 | 0.32 |  |
| ENVINS3 | Hispanic | | 24591 | -1.15 | -0.83 | 0.32 |  |
| ENVMEN3 | Asian | | 3886 | -0.16 | 0.16 | 0.32 |  |
| SAFBUL3 | White | | 54877 | -0.45 | -0.12 | 0.33 |  |
| SAFPSF4 | Hispanic | | 12826 | -1.28 | -0.94 | 0.33 |  |
| SAFBUL3 | Other | | 3742 | -0.45 | -0.11 | 0.34 |  |
| ENVMEN1 | Afr./Black | | 4675 | -0.08 | 0.26 | 0.34 |  |
| SAFBUL1 | Afr./Black | | 13102 | -0.94 | -0.59 | 0.35 |  |
| ENVMEN6 | Afr./Black | | 3349 | -0.39 | -0.04 | 0.35 |  |
| SAFEMO7 | Asian | | 2129 | -0.36 | 0 | 0.36 |  |
| SAFEMO7 | Hispanic | | 8042 | -0.36 | 0.01 | 0.37 |  |
| SAFBUL3 | Asian | | 6022 | -0.45 | -0.07 | 0.38 |  |
| SAFBUL12 | Hispanic | | 24593 | 0.36 | 0.74 | 0.38 |  |
| ENVMEN4 | Asian | | 3896 | -0.03 | 0.36 | 0.39 |  |
| ENGPAR3 | Hispanic | | 8014 | -0.77 | -0.37 | 0.4 |  |
| SAFPSF7 | Afr./Black | | 9116 | 0.93 | 1.33 | 0.40 |  |
| ENVMEN1 | Hispanic | | 12507 | -0.08 | 0.32 | 0.40 |  |
| SAFBUL7 | Afr./Black | | 4410 | 0.82 | 1.24 | 0.42 |  |
| SAFEMO7 | Afr./Black | | 3382 | -0.36 | 0.06 | 0.42 |  |
| SAFPSF5 | Hispanic | | 7973 | -0.96 | -0.52 | 0.44 |  |
| ENVINS5 | Asian | | 6014 | -0.64 | -0.16 | 0.48 |  |
| ENVMEN6 | Asian | | 2121 | -0.39 | 0.11 | 0.50 |  |
| SAFPSF1 | Hispanic | | 24669 | -0.86 | -0.33 | 0.53 |  |
| ENGCLC3 | Asian | | 3891 | 0.44 | 0.98 | 0.54 |  |
| SAFPSF1 | Afr./Black | | 9123 | -0.86 | -0.32 | 0.54 |  |
| SAFBUL11 | Hispanic | | 7940 | -1.37 | -0.77 | 0.60 | 0.48 |
| SAFBUL11 | Other | | 1442 | -1.37 | -0.61 | **0.75** | **0.74** |
| SAFBUL11 | Afr./Black | | 3320 | -1.37 | -0.61 | **0.76** | 0.65 |
| SAFBUL11 | Asian | | 2108 | -1.37 | -0.24 | **1.13** | **1.08** |

1DIF analyses examine if item has the same difficulty as its average difficulty for all groups; 2New DL items bolded

**Appendix I3, (Generalizability): DIF by Low-Income Status1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item2** | **Not Low-Income measure** | **Low-Income measure** | **VOCAL DIF Size** | **Dimension DIF Size** |
| SAFBUL11 | -1.68 | -1.01 | -0.67 | 0.55 |
| SAFBUL12 | 0.09 | 0.65 | -0.56 |  |
| SAFPSF5 | -1.21 | -0.66 | -0.55 |  |
| ENGPAR3 | -1 | -0.47 | -0.53 |  |
| ENGCLC4 | -0.41 | 0.09 | -0.49 |  |
| SAFPSF4 | -1.5 | -1.03 | -0.46 |  |
| SAFBUL7 | 0.61 | 1.05 | -0.44 |  |
| SAFBUL2 | 0.46 | 0.88 | -0.43 |  |
| ENGPAR1 | -0.23 | 0.2 | -0.42 |  |
| SAFBUL5 | -0.58 | -0.16 | -0.42 |  |
| SAFEMO6 | -1.13 | -0.73 | -0.4 |  |
| SAFPSF7 | 0.76 | 1.12 | -0.36 |  |
| ENGPAR2 | -0.75 | -0.4 | -0.35 |  |
| SAFBUL13 | 0.5 | 0.83 | -0.33 |  |
| SAFBUL15 | -0.07 | 0.26 | -0.33 |  |
| SAFPSF1 | -0.74 | -0.41 | -0.33 |  |
| ENGCLC1 | -1.11 | -0.79 | -0.32 |  |
| ENVMEN1 | -0.06 | 0.26 | -0.32 |  |
| SAFPSF3 | 0.4 | 0.72 | -0.32 |  |
| SAFBUL10 | -1.53 | -1.23 | -0.3 |  |
| ENVMEN2 | 0.04 | 0.29 | -0.25 |  |
| SAFBUL6 | -1.31 | -1.06 | -0.25 |  |
| SAFEMO7 | -0.27 | -0.02 | -0.25 |  |
| ENGCLC2 | -1.33 | -1.09 | -0.24 |  |
| SAFBUL14 | 0.18 | 0.42 | -0.24 |  |
| ENVDIS2 | -0.33 | -0.11 | -0.23 |  |
| SAFEMO3 | -0.23 | -0.01 | -0.22 |  |
| **BETTER CONTRASTING ITEMS ARE OMITTED** | | | | |
| ENVDIS8 | 0.65 | 0.43 | 0.22 |  |
| ENVINS14 | 1.84 | 1.62 | 0.22 |  |
| **ENGPAR16** | 0.88 | 0.65 | 0.23 |  |
| ENVINS13 | 0.57 | 0.35 | 0.23 |  |
| ENGPAR5 | 0.49 | 0.25 | 0.24 |  |
| ENGPAR6 | -0.32 | -0.58 | 0.26 |  |
| **ENGPAR24** | 1.3 | 1.01 | 0.29 |  |
| **ENVINS20** | 0.72 | 0.42 | 0.3 |  |
| ENGPAR12 | 0.68 | 0.37 | 0.31 |  |
| **ENVINS23** | -0.37 | -0.68 | 0.31 |  |
| **ENGPAR25** | 0.83 | 0.48 | 0.35 |  |
| ENGPAR10 | 0.93 | 0.57 | 0.36 |  |
| **ENVINS21** | 0.28 | -0.09 | 0.37 |  |
| ENGPAR9 | -0.06 | -0.44 | 0.38 |  |
| ENGPAR8 | 0.74 | 0.28 | 0.46 |  |
| ENGPAR7 | 0.83 | 0.36 | 0.48 |  |
| ENGPAR14 | 1.83 | 1.26 | 0.57 |  |
| **ENVINS18** | 0.44 | -0.2 | 0.64 |  |

1This DIF analyses examines if the item has the same difficulty for the two comparison groups; 2New DL items bolded

**Appendix I4, (Generalizability): DIF by English learner (EL) Status1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item2** | **Not EL measure** | **EL measure** | **VOCAL**  **DIF Size** | **Omitted from calibration** | **Dimension DIF Size** |
| SAFBUL11 | -1.51 | 0.11 | **-1.62** | G10, Yes | **-1.27** |
| SAFBUL10 | -1.51 | -0.17 | **-1.34** | G10, Yes | **-0.99** |
| SAFPSF5 | -1.05 | 0.08 | **-1.13** | G10, Yes | **-0.78** |
| SAFPSF4 | -1.37 | -0.35 | **-1.02** | G8, No | **-0.84** |
| SAFBUL16 | -0.21 | 0.69 | **-0.9** | G10, Yes | -0.53 |
| SAFEMO6 | -1.01 | -0.11 | **-0.9** | G10, Yes | -0.55 |
| SAFBUL12 | 0.26 | 1.06 | **-0.8** | G4,5 No | -0.62 |
| SAFBUL5 | -0.44 | 0.3 | **-0.75** | G8, No | -0.55 |
| ENGPAR3 | -0.82 | -0.08 | **-0.74** | G10, Yes | **-0.85** |
| SAFBUL15 | 0.02 | 0.75 | **-0.73** | G8, No | -0.52 |
| ENGPAR1 | -0.12 | 0.6 | **-0.72** | G4,5 No | **-0.84** |
| SAFBUL13 | 0.59 | 1.23 | -0.64 |  | -0.68 |
| ENGCLC4 | -0.23 | 0.35 | -0.58 |  |  |
| SAFPSF8 | -0.21 | 0.33 | -0.54 |  |  |
| ENGCLC1 | -1 | -0.5 | -0.51 |  |  |
| **ENGCLC10** | -1.02 | -0.52 | -0.5 |  |  |
| ENGREL3 | -1.45 | -0.95 | -0.5 |  |  |
| SAFPSF7 | 0.87 | 1.37 | -0.5 |  |  |
| SAFPSF3 | 0.51 | 0.99 | -0.48 |  |  |
| SAFBUL6 | -1.26 | -0.79 | -0.47 |  |  |
| ENGCLC7 | -1.07 | -0.61 | -0.46 |  |  |
| SAFBUL7 | 0.77 | 1.16 | -0.39 |  |  |
| SAFBUL8 | -0.5 | -0.13 | -0.37 |  |  |
| SAFPSF1 | -0.63 | -0.27 | -0.35 |  |  |
| ENGCLC2 | -1.26 | -0.92 | -0.34 |  |  |
| SAFBUL14 | 0.26 | 0.59 | -0.33 |  |  |
| **BETTER CONTRASTING ITEMS ARE OMITTED** | | | | | |
| ENGPAR16 | 0.81 | 0.47 | 0.33 |  |  |
| ENGPAR4 | 0.92 | 0.57 | 0.34 |  |  |
| SAFEMO9 | 0.71 | 0.38 | 0.34 |  |  |
| ENVINS13 | 0.5 | 0.14 | 0.36 |  |  |
| ENGPAR8 | 0.58 | 0.19 | 0.39 |  |  |
| ENVDIS1 | 1.64 | 1.24 | 0.41 |  |  |
| ENVDIS7 | 0.49 | 0.03 | 0.46 |  |  |
| **ENGPAR25** | 0.73 | 0.25 | 0.48 |  |  |
| ENVDIS8 | 0.59 | 0.07 | 0.52 |  |  |
| ENGPAR10 | 0.81 | 0.27 | 0.54 |  |  |
| **ENVINS21** | 0.16 | -0.38 | 0.54 |  |  |
| **ENVINS18** | 0.21 | -0.34 | 0.55 |  |  |
| ENVINS14 | 1.8 | 1.24 | 0.57 |  |  |
| **ENVINS20** | 0.63 | 0.06 | 0.57 |  |  |
| ENGPAR12 | 0.59 | -0.03 | 0.62 | G10, Yes | 0.54 |
| ENGPAR7 | 0.69 | 0.02 | 0.67 | G4,5 No | 0.56 |
| **ENGPAR24** | 1.22 | 0.53 | 0.69 | G10, New | 0.58 |
| ENVINS12 | 0.79 | 0.01 | **0.78** | G10, Yes | 0.36 |
| ENGPAR14 | 1.65 | 0.61 | **1.04** | G10, Yes | **0.75** |

1This DIF analyses examines if the item has the same difficulty for the two comparison groups;2New DL items bolded;3Not applicable

**Appendix I5, (Generalizability): DIF by Students with disabilities (SWD) Status1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item2** | **Not SWD measure** | **SWD measure** | **VOCAL**  **DIF Size** | **DIMENSION**  **DIF Size** |
| DENGCLC4 | -0.4 | 0.76 | **-1.17** | **-1.10** |
| ENGPAR3 | -0.85 | -0.35 | -0.5 |  |
| ENGPAR1 | -0.12 | 0.36 | -0.48 |  |
| SAFBUL10 | -1.48 | -1.02 | -0.46 |  |
| **ENGCLC10** | -1.06 | -0.66 | -0.39 |  |
| SAFEMO6 | -1.01 | -0.64 | -0.37 |  |
| SAFBUL16 | -0.21 | 0.15 | -0.36 |  |
| SAFBUL6 | -1.26 | -0.91 | -0.34 |  |
| SAFBUL12 | 0.3 | 0.63 | -0.34 |  |
| SAFPSF5 | -1.02 | -0.68 | -0.34 |  |
| SAFBUL5 | -0.44 | -0.11 | -0.33 |  |
| ENGCLC3 | 0.39 | 0.68 | -0.28 |  |
| SAFBUL15 | 0.03 | 0.31 | -0.28 |  |
| ENVINS5 | -0.5 | -0.23 | -0.27 |  |
| SAFBUL11 | -1.42 | -1.15 | -0.26 |  |
| **ENGPAR22** | -0.07 | 0.18 | -0.25 |  |
| SAFBUL7 | 0.78 | 1.01 | -0.24 |  |
| ENGCLC7 | -1.07 | -0.84 | -0.23 |  |
| SAFBUL2 | 0.62 | 0.85 | -0.23 |  |
| SAFBUL14 | 0.25 | 0.47 | -0.23 |  |
| SAFPSF4 | -1.32 | -1.09 | -0.23 |  |
| SAFBUL13 | 0.61 | 0.8 | -0.19 |  |
| ENGCLC2 | -1.25 | -1.07 | -0.17 |  |
| ENVDIS2 | -0.26 | -0.08 | -0.17 |  |
| **BETTER CONTRASTING ITEMS ARE OMITTED** | | | | |
| ENVDIS7 | 0.48 | 0.32 | 0.16 |  |
| ENVINS11 | -0.3 | -0.47 | 0.17 |  |
| SAFEMO13 | 1.53 | 1.34 | 0.18 |  |
| SAFEMO8 | 0.81 | 0.61 | 0.2 |  |
| ENVDIS6 | 0.25 | 0.05 | 0.2 |  |
| ENVMEN4 | 0.11 | -0.08 | 0.2 |  |
| ENVDIS8 | 0.59 | 0.38 | 0.21 |  |
| ENVMEN9 | 1.09 | 0.84 | 0.24 |  |
| **ENGPAR25** | 0.72 | 0.44 | 0.28 |  |
| **ENVINS21** | 0.17 | -0.11 | 0.28 |  |
| ENVINS2 | -1 | -1.29 | 0.29 |  |
| **ENVINS20** | 0.65 | 0.34 | 0.3 |  |
| **ENGPAR24** | 1.23 | 0.92 | 0.31 |  |
| ENVMEN6 | -0.08 | -0.39 | 0.31 |  |
| ENGPAR7 | 0.68 | 0.36 | 0.32 |  |
| SAFEMO4 | 0.8 | 0.45 | 0.35 |  |
| ENVINS18 | 0.23 | -0.16 | 0.39 |  |
| ENGPAR14 | 1.65 | 1.22 | 0.43 |  |

1This DIF analyses examines if the item has the same difficulty for the two comparison groups;2New DL items bolded;3Not applicable

**Appendix J1, (External Validity): School-level correlations: Behavior indicators and VOCAL dimension scores (all grades)1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Indicator (Number of schools)** | **Overall 7** | **Engagement 7** | **Safety7** | **Environment7** |
| Attendance rate2 (N = 1277) | 0.21\*\* | 0.17\*\* | 0.30\*\* | 0.10\*\* |
| Chronically absent3 (N = 1277) | -0.23\*\* | -0.16\*\* | -0.37\*\* | -0.12\*\* |
| Discipline rate4 (N = 1163) | -0.58\*\* | -0.54\*\* | -0.60\*\* | -0.51\*\* |
| In-school suspension (N = 682)5 | -0.29\*\* | -0.28\*\* | -0.29\*\* | -0.25\*\* |
| Out-school suspension (N = 682)6 | -0.39\*\* | -0.33\*\* | -0.47\*\* | -0.28\*\* |

1Data based on schools with greater than or equal to 10 students contributing to the aggregate VOCAL score and includes all of the potential four grades enrolled within the school; 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; 7\*\**p*<0.01; \* *p*<0.05; NSNot Significant.

**Appendix J2, (External Validity): School-level correlations by grade: Academic achievement or growth scores and overall VOCAL scores1,2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade** | **Grade 4**  **(N = 746)** | **Grade 5**  **(N = 690)** | **Grade 8**  **(N = 405)** | **Grade 10**  **(N = 253)** |
| English Language Arts and Literacy scaled score | 0.30\*\* | 0.27\*\* | 0.27\*\* | 0.03NS |
| ELA and Literacy student growth percentile | 0.18\*\* | 0.22\*\* | 0.27\*\* | 0.07 NS |
| Mathematics scaled score | 0.34\*\* | 0.27\*\* | 0.26\*\* | 0.04 NS |
| Mathematics student growth percentile | 0.21\*\* | 0.17\*\* | 0.18\*\* | 0.07 NS |

1Data based on schools with greater than 10 students contributing to the aggregate VOCAL score; 2\*\**p*<0.01; \**p*<0.05; NS: not significant.

**Appendix J3, (External Validity): School-level correlations by grade: English Language Arts (ELA) academic achievement and VOCAL dimension scores1,2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade (Number)** | **Overall** | **Engagement** | **Safety** | **Environment** |
| Grade 4 (N = 746) 7 | 0.30\*\* | 0.04NS | 0.47\*\* | 0.14\*\* |
| Grade 5 (N = 690) | 0.27\*\* | 0.10\*\* | 0.45\*\* | 0.11\*\* |
| Grade 8 (N = 405) | 0.27\*\* | 0.14\*\* | 0.49\*\* | -0.01 NS |
| Grade 10 (N = 253) | 0.03 NS | 0.03 NS | 0.24\*\* | -0.23\*\* |

1Data based on schools with greater than 10 students contributing to the aggregate VOCAL score; 2\*\**p*<0.01; \**p*<0.05; NS: not significant.

**Appendix J4, (External Validity): School-level correlations by grade: English Language Arts (ELA) academic growth and VOCAL dimension scores1,2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade (Number)** | **Overall** | **Engagement** | **Safety** | **Environment** |
| Grade 4 (N = 746) 7 | 0.18\*\* | 0.02NS | 0.28\*\* | 0.06 NS |
| Grade 5 (N = 690) | 0.22\*\* | 0.14\*\* | 0.30\*\* | 0.13\*\* |
| Grade 8 (N = 405) | 0.27\*\* | 0.20\*\* | 0.33\*\* | 0.16 NS |
| Grade 10 (N = 253) | 0.07 NS | 0.06 NS | 0.15\*\* | -0.10 NS |

1Data based on schools with greater than 10 students contributing to the aggregate VOCAL score; 2\*\**p*<0.01; \**p*<0.05; NS: not significant.

**Appendix J5, (External Validity): School-level correlations by grade: Mathematics academic achievement and VOCAL dimension scores1,2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade (Number)** | **Overall** | **Engagement** | **Safety** | **Environment** |
| Grade 4 (N = 746) 7 | 0.34\*\* | 0.09\* | 0.50\*\* | 0.17\*\* |
| Grade 5 (N = 690) | 0.27\*\* | 0.10\*\* | 0.45\*\* | 0.11\*\* |
| Grade 8 (N = 405) | 0.26\*\* | 0.15\*\* | 0.46\*\* | -0.005NS |
| Grade 10 (N = 253) | 0.04 NS | 0.05 NS | 0.25\*\* | -0.23\*\* |

1Data based on schools with greater than 10 students contributing to the aggregate VOCAL score 2\*\**p*<0.01; \**p*<0.05; NS: not significant.

**Appendix J6, (External Validity): School-level correlations by grade: Mathematics academic growth and VOCAL dimension scores1,2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Grade (Number)** | **Overall** | **Engagement** | **Safety** | **Environment** |
| Grade 4 (N = 746) | 0.21\*\* | 0.09\* | 0.28\*\* | 0.10\*\* |
| Grade 5 (N = 690) | 0.17\*\* | 0.11\*\* | 0.22\*\* | 0.10\*\* |
| Grade 8 (N = 405) | 0.18\*\* | 0.13\*\* | 0.23\*\* | 0.08 NS |
| Grade 10 (N = 253) | 0.07 NS | 0.03 NS | 0.19\*\* | -0.10 NS |

1Data based on schools with greater than 10 students contributing to the aggregate VOCAL score; 2\*\**p*<0.01; \**p*<0.05; NS: not significant.