Reliability and Validity of Panorama's Survey Topics for Students: 2024 Update with Addendum for Versioned Content

Table of Contents

Executive Summary

Panorama Education's student surveys cover a range of topics about students' life skills and school environments. Students are more than just respondents for these self-report surveys: By reflecting on their experiences and sharing their perspectives, they are active agents in their own learning and in their school's improvement. We developed our student surveys using a rigorous six-step design process that builds reliability and validity into them from the beginning (Gehlbach & Brinkworth, 2011; see also Artino, La Rochelle, DeZee, & Gehlbach, 2014). We analyzed data from millions of students in thousands of K-12 schools across the U.S. to demonstrate the psychometric strength of our surveys. Specifically, findings showed that:

1. Almost all students answered all of the questions in a topic. Even for the topic with the lowest completion rate, 93.3% of respondents answered every question, suggesting that non-response bias is not an issue for our survey results.
2. Our survey topics exhibit strong reliability by exceeding the conventional thresholds of internal consistency.
3. Our topics demonstrate strong structural validity, meaning that the items making up each topic do indeed belong together as part of a single construct.
4. Our topics evidence both convergent and discriminant validity in that they show higher correlations with theoretically related constructs and lower correlations with theoretically unrelated constructs.

Panorama's surveys are available to all on our website. These measures of students' life skills and school environments offer a research-backed means to inform the work of educators, policymakers, and researchers.

Introduction

In this report, we describe the development of Panorama's student survey topics and provide updated evidence for their reliability and validity. We consolidate and expand on past reports by covering the full range of topics from our student surveys, including those that measure life skills, as well as those that capture students' perceptions of their school, classrooms, and teachers. This report presents psychometric results from much larger datasets than past reports and also takes advantage of a broader range of data sources. Our goal in presenting these results is to inform not just educators and experts from the schools, districts, states, and organizations that we partner with, but also the larger community of policymakers and practitioners interested in educational measurement, human development, and school improvement.

For educators interested in measuring life skills or climate variables for these purposes, self-report surveys possess principled and pragmatic advantages over other measurement strategies. They are uniquely efficient and cost-effective, allowing educators to quickly and affordably census their entire student body on an array of topics. Unlike other methods (e.g., classroom observation, performance tasks) that tend to treat students primarily as data sources, climate and life skills surveys typically treat students as active agents of school improvement and of their own learning; in this sense, they can serve as both measurement instruments and the basis for educational interventions. Finally, for educators primarily interested in students' internal mental states (e.g., perceptions of safety, feelings of belonging, judgments of self-efficacy), self-report is a natural fit over other methods, such as teacher-report, that require others to infer students' thoughts or feelings. These and other benefits (e.g., ease of reporting, standardization, understandability of results) explain the prevalence of self-report surveys in climate and life skills measurement.

In considering the quality of these or any other measurement instrument, bedrock principles of psychometrics are worth keeping front of mind. First, despite established criteria of what constitutes "good" for many statistics, reliability and validity exist in shades of gray. When we claim here or elsewhere that our survey instruments are "valid and reliable," we mean that they meet established standards of educational measurement, not that reliability or validity are black-and-white determinations. Second, reliability and validity depend on use; they are not fixed properties of a measurement instrument (Messick, 1995a). If schools are funded, if teachers are compensated, or if students' grades are based on self-report data, then these contingencies will likely contaminate the measurement process. For this reason, the results we present here are for the informative applications that motivate our client partnerships.

Methods

As described below, we developed a broad range of student measures by following best practices in the science of survey development. Following up initial validation efforts, we used a

large, de-identified dataset from Panorama's partners (with their permission) to comprehensively examine these topics' psychometric properties.

Topic Development

Education researchers and practitioners at the Harvard University Graduate School of Education developed Panorama's student surveys, which came out of the six-step design process by Gehlbach and Brinkworth (2011; see also Artino et al., 2014).[1] An outline of the process is below. To the best of our knowledge, this process is unsurpassed in terms of its rigor and capacity to minimize survey error. The strengths of this process come from two approaches.

First, this process builds evidence of validity—specifically, content validity and substantive validity (Messick, 1995b)—into each survey topic from the outset of the design process. The six key steps in the process include literature review, interviews and focus groups, synthesis of indicators, item (question) creation, expert review, and cognitive pre-testing and interviewing. Upon completion of these six steps and a round of revisions to the items, the topics were subjected to large-scale pilot tests.

Second, this process ensures that each survey item is designed to adhere to the science of survey design's best practices (Gehlbach & Artino, 2018; Dillman, Smyth, & Christian, 2014; Fowler, 2013). For example, researchers have concluded that designing survey items as statements, particularly ones that require respondents to agree or disagree, is likely to inject additional measurement error and bias into responses.

Numerous surveys used by educators unfortunately fail to adhere to these well-established survey design practices. For example, asking questions with response options that are linked to the underlying concept is the preferred practice (Dillman et al., 2014; Krosnick, 1999a; Saris, Revilla, Krosnick, & Shaeffer, 2010)—e.g., rather than posing statements that respondents are supposed to agree or disagree with. Failing to label all response options, using numeric rather than verbal labels, and using too few response options are other commonly violated best practices (Artino et al., 2014; Dillman et al., 2014; Krosnick, 1999b; Weng, 2004). As a survey topic violates more of these best practices, the amount of measurement error and bias grows. The topics that comprise Panorama's survey instruments adhere to these best practices, which was confirmed during the expert review step.

Table 1 summarizes Panorama's student survey topics, including each construct's definition and recommended key scholarly references. All of the survey content is available on [our website](#).

---

[1] To bring non-Panorama measures of social awareness and self-management to schools and districts upon request while also keeping consistency and Panorama's high standards for research-backed surveys, our team adapted these topics to better align with best practices in survey design. For example, agree/disagree statements were rephrased as questions.

PANORAMA
EDUCATION

Table 1. Panorama Student Survey Topics		
Construct/Topic	**Definition**	**Key Reference(s)**
Classroom Effort	How much effort students put into school and learning	Covington (2000); Ryan & Deci (2000)
Climate	Perceptions of the overall social and learning climate of the school	Thapa, Cohen, Guffey, & Higgins-D'Alessandro (2013)
Emotion Regulation	How well students regulate their emotions	Duckworth, Gendler, & Gross (2014)
Engagement	How attentive and invested students are in school	Fredricks, Blumenfeld, & Paris (2004)
Grit	How well students are able to persevere through setbacks to achieve important long-term goals	Duckworth & Gross (2014)
Growth Mindset	Student perceptions of whether they have the potential to change those factors that are central to their performance in school	Dweck (2008); Yeager et al. (2019)
Learning Strategies	How well students deliberately use strategies to manage their own learning processes generally	Fiorella & Mayer (2016); Pintrich & De Groot (1990)
Pedagogical Effectiveness	Perceptions of the quality of teaching and amount of learning students experience from a particular teacher	Cantrell & Kane (2013); Good (2014)
Rigorous Expectations	How much students feel that their teachers hold them to high expectations around effort, understanding, persistence and performance in class	Ames (1992); Jussim & Harber (2005)
School Safety	Student perceptions of physical and psychological safety while at school	Cornel & Mayer (2010)
Self-Efficacy	How much students believe they can succeed in achieving academic outcomes	Usher & Pajares (2008)
Self-Management	How well students manage their emotions, thoughts, and behaviors in different situations	Transforming Education (2016)
Sense of Belonging	How much students feel that they are valued members of the school community	Cohen & Garcia (2008); Osterman (2000)
Social Awareness	How well students consider the perspectives of others and empathize with them	Transforming Education (2016)
Teacher-Student Relationships	How strong the social connection is between teachers and students within and beyond the school	Pianta, Hamre, & Allen (2012); Roorda, Koomen, Spilt, & Oort (2011)

Some survey topics have parallel classroom and school forms (e.g., Classroom Climate and School Climate, respectively) to help educators target feedback to classroom and schools, respectively. Additionally, certain topics have parallel forms based on students' grade level (3-5th grade vs 6-12th grade) to accommodate differences in reading comprehension or educational context (i.e., elementary vs secondary schooling). Since the differences between parallel forms, when they exist, are relatively minor (e.g.,"How sure are you..." vs "How confident are you...") we do not present separate analyses in this report for each form.

Data Source

For the analyses we report here, we relied on survey data collected during the 2016-17 school year from approximately 3,500,000 students in over 5,900 schools. We present further information on this sample in the Results section below.

Scoring

All survey items have five fully-labeled verbal response options, except for a few bipolar questions with seven response options. To calculate topic scores, we converted each of the five response options to a 1-5 integer value (with 5 always reflecting a more positive response) and averaged across item scores within the topics. For items with 7 response options, we arithmetically transformed those responses to a 1-5 scale before taking the average. Unless otherwise noted, we excluded any student who skipped more than half of the items in a topic.

Results

Sample Demographics

Table 2 presents the demographics for the student sample analyzed in this report.[2] The sample shows considerable demographic diversity that either meets or exceeds that found nationally (McFarland et al., 2019). As expected, it contains an even distribution of students with respect to gender and grade level. Reflecting the schools and districts partnering with Panorama, the sample contains relatively more underrepresented populations, English Language Learners (ELLs) and students eligible for free or reduced-priced lunch (FRPL) relative to the national student population.

---

[2] Since some students were missing demographic data—due to survey non-response, schools or districts electing to not include demographic data with their surveys, or certain demographics not being recorded in school and district data systems—percentages reported reflect just those students with non-missing data for the given area. The percentage of students with missing demographics ranged from a low of 5.2% for grade level to a high of 37.4% for free/reduced-price lunch eligibility. Note that the percentages may not total to 100.0 due to rounding error.

Table 2. Student Demographics	
Demographic Group	**%**
Female	49.6
Male	50.4
Asian	5.8
Black	13.3
Latinx	53.0
Multiracial	1.9
Native American	0.5
Pacific Islander	1.6
White	23.9
3rd Grade	9.5
4th Grade	11.0
5th Grade	11.6
6th Grade	10.9
7th Grade	11.5
8th Grade	10.7
9th Grade	10.4
10th Grade	9.1
11th Grade	8.3
12th Grade	7.1
English language learner	21.3
Eligible for free/reduced-price lunch	76.6

Survey Satisficing

Satisficing (a portmanteau of *satisfy* and *suffice*) is a decision-making strategy in which individuals expend just enough mental effort to reach an acceptable (as opposed to optimal) solution (Simon, 1957). In taking surveys, respondents can exhibit satisficing in various forms—e.g., by skipping items, selecting a "don't know" response option, or giving identical responses across multiple

items (called straightlining)—all of which represent potential threats to a topic's reliability or validity (see Barge & Gehlbach, 2012; Krosnick, 1991).

Past large-sample research on student life skills surveys administered by Panorama (n = 409,721) has shown that satisficing is a relatively minor concern (Vriesema & Gehlbach, 2019). To confirm that conclusion, we investigated the extent of satisficing in our entire 2016-17 survey dataset by examining the combined effects of early termination (i.e., submitting an unfinished survey) and item skipping (i.e., neglecting to answer individual survey questions). Both forms of satisficing manifest as missing data and could potentially bias results if respondents with missing data differ from those with complete data on the constructs of interest (e.g., Holt & Elliot, 1991).

The vast majority of respondents answered all survey questions within each topic: On average, 95.7% of respondents answered all questions from a topic. Even the topic with the lowest full-completion rate (also the topic with the most items), Self-Management, still had 93.3% of respondents completing every question, leaving only 6.7% of respondents skipping one or more questions. An item-level analysis revealed similarly low levels of satisficing: The mean missing data rate across all items was only 1.2%, and the max rate of missing data across all items was 1.8%. Figure 1 presents the distribution of the percentage of missing data across all survey items. Taken together, these results affirm that early survey termination or item skipping do not threaten the quality of the data schools collect through Panorama.[3]
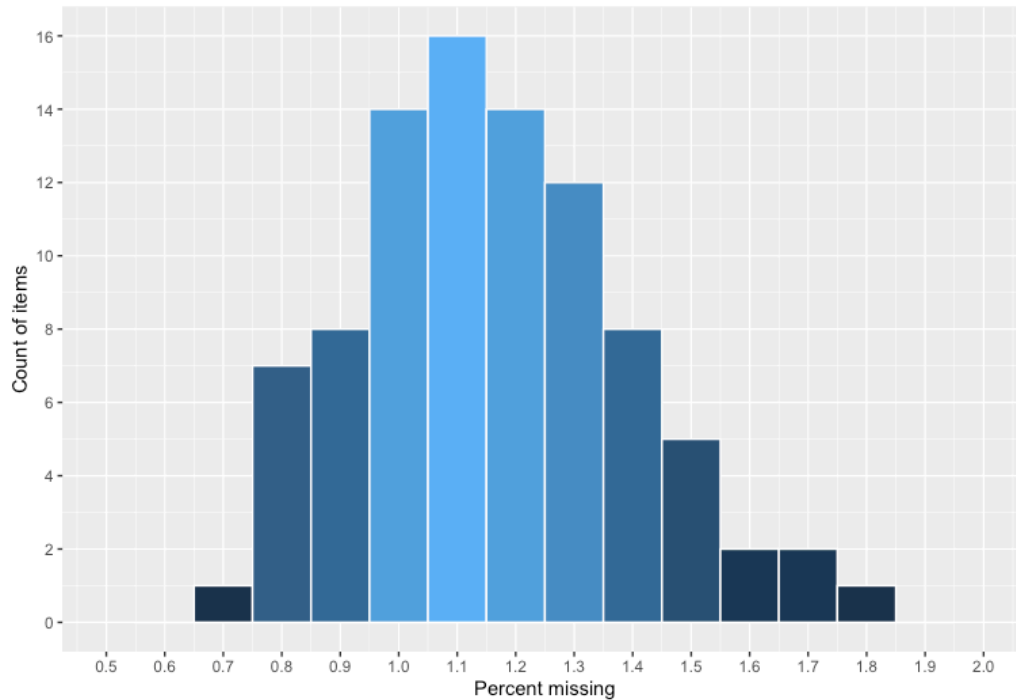
Table 3. Full Completion Rates by Topic	
Topic	% Students with Complete Data
Climate	96.6
Self-Efficacy	95.9
Classroom Effort	97.5
Engagement	95.4
Emotion Regulation	96.6
Growth Mindset	95.3
Grit	95.7
Learning Strategies	95.4
Pedagogical Effectiveness	93.6
Rigorous Expectations	95.8
Social Awareness	94.7
Sense of Belonging	96.7

---

[3] For satisficing analyses, we included all available data; unlike other analyses, we did not exclude data from participants who did not answer at least half the questions from a topic.

Self-Management	93.3
School Safety	95.8
Teacher-Student Relationships	96.8

Figure 1. Distribution of Missing Data across All Survey Items

Reliability

Reliability, as assessed through Cronbach's alpha, is essentially a measure of signal-to-noise (DeVellis, 2016) with higher values reflecting more "signal" and less "noise." Put differently, Cronbach's alpha measures how similarly students respond to items from the same topic, i.e., a topic's internal consistency (Streiner, 2003). As shown in Table 4, all of Panorama's survey topics met or exceeded the typical sufficiency threshold of .70.

Structural Validity

We conducted confirmatory factor analyses to assess the structural validity of each survey topic (see Messick, 1995b). More specifically, we examined whether each topic had the statistical structure we expected, i.e., whether it measures only a single dimension (and not multiple dimensions). Table 4 presents the two key statistics from each analysis: the comparative fit index (CFI) and the root mean square error of approximation (RMSEA).

With a maximum possible value of 1, CFI measures how well the data from each topic fit a one-dimensional solution. Historically, a CFI of 0.90 or greater has been considered sufficient,

though a value closer to 0.95 or above is preferred. As shown in Table 4, all but one topic meet the preferred threshold, with many effectively at the upper bound of 1. RMSEA is a complementary measure of model fit, with lower values indicating better fit. All but one topic were at or under the typical threshold of 0.08. The same topic, Self-Management, missed the conventional thresholds for both statistics, but only marginally; notably, it is also the topic with the largest number of questions. (For a discussion of the CFI and RMSEA, including thresholds to determine model fit, see Hu & Bentler, 1999; Laverdière, Morin, & St-Hilaire, 2013.)

Table 4. Reliability and Structural Validity Results by Topic			
Topic	**α**	**CFI**	**RMSEA**
Climate	0.84	1.00	0.04
Self-Efficacy	0.78	1.00	0.03
Classroom Effort	0.75	1.00	0.03
Engagement	0.87	1.00	0.05
Emotion Regulation	0.84	0.99	0.05
Growth Mindset	0.79	0.96	0.08
Grit	0.74	0.99	0.05
Learning Strategies	0.83	0.99	0.05
Pedagogical Effectiveness	0.94	0.99	0.06
Rigorous Expectations	0.86	0.99	0.06
Social Awareness	0.80	0.96	0.06
Sense of Belonging	0.85	0.99	0.06
Self-Management	0.83	0.89	0.09
School Safety	0.73	0.95	0.07
Teacher-Student Relationships	0.90	0.99	0.07
Note: α = Cronbach's alpha, CFI = comparative fit index, RMSEA = root mean square error of approximation; all results based on 2016-17 sample and rounded to two decimal places.			

Convergent and Discriminant Validity

Convergent and discriminant validity assess how much a measurement instrument, in this case a survey topic, measures what it is designed to measure. An instrument demonstrates convergent validity when it correlates as expected with measures of theoretically similar constructs (or alternative measures of the same construct). Conversely, an instrument demonstrates

discriminant validity when it correlates minimally with measures of theoretically-dissimilar constructs. If students' responses on our Sense of Belonging topic correlated positively with their responses on our Teacher-Student Relationships topic, for example, that would be evidence of convergent validity; and if students' Sense of Belonging responses showed a relatively small correlation with their Classroom Effort responses or standardized test scores, that would be evidence of discriminant validity (for more, see Messick, 1995a).

To address convergent and discriminant validity, we computed intercorrelations among all student topics at the student and school level and compared those with what one would expect based on extant scholarship. As shown in Figure 2, the survey topics correlate with each other largely as expected. With only one exception (a trivially negative school-level correlation between Social Awareness and Engagement), the correlations are all positive, with stronger correlations for more related constructs (e.g., Sense of Belonging and Teacher-Student Relationships) and weaker correlations for less related constructs (e.g., School Safety and Growth Mindset).

As one of many possible examples of convergent validity, we see particularly strong correlations (at both the student and school level) between our Teacher-Student Relationships, Rigorous Expectations, Sense of Belonging, and Climate topics. Based on prior work showing the centrality of students' relationships with their teachers to their social experience of school more generally—as well as the impact of teachers' setting high expectations for their students—we would expect these topics to be highly related (see Anderman, 2003; Brinkworth, McIntyre, Juraschek, & Gehlbach, 2018; Lee, 2012)

As an example of discriminant validity, we see relatively no meaningful correlations between students' perception of the safety of their environment as measured by our School Safety topic and their life skills competencies (e.g., Social Awareness, Growth Mindset), perceptions of their teachers (e.g., Pedagogical Effectiveness, Rigorous Expectations), or even feeling that they belong at school (e.g., Sense of Belonging). Based on prior psychometric evidence and research (e.g., Skiba et al., 2004; Thapa et al., 2013) , one would expect students' safety experience to diverge from these other variables.

Figure 2. Topic Intercorrelations

	Climate	Self-Efficacy	Classroom Effort	Engagement	Emotion Regulation	Growth Mindset	Grit	Learning Strategies	Pedagogical Eff	Rigorous Expectations	Social Awareness	Sense of Belonging	Self-Management	School Safety	Teacher-Student Rel
Climate	—			0.7		0.5	0.3	0.7	0.6	0.6		0.7		0.4	0.9
Self-Efficacy	0.4	—				0.7	0.2				0.7	0.4	0.7	-0.1	0.4
Classroom Effort	0.3	0.4	—												
Engagement	0.7	0.4	0.4	—		0.6	0.5	0.8	0.6	0.6	-0.1	0.6	0.0	0.3	0.7
Emotion Regulation		0.4	0.4	0.3	—	0.1	0.1					0.2			
Growth Mindset	0.3	0.3	0.2	0.4	0.2	—	0.3	0.5		0.6	0.6	0.5	0.7	0.4	0.5
Grit	0.4	0.5	0.5	0.4	0.4	0.3	—	0.6	0.3	0.3	0.8	0.7	0.3	0.0	0.4
Learning Strategies	0.5	0.6	0.5	0.6	0.4	0.4	0.7	—		0.4		0.8		0.5	0.7
Pedagogical Effectiveness	0.8		0.5	0.8		0.4	0.5	0.6	—	0.8		0.5		0.0	0.9
Rigorous Expectations	0.7	0.5	0.4	0.7		0.3	0.4	0.6	0.8	—		0.7		0.3	0.9
Social Awareness		0.5	0.5	0.5	0.5	0.3	0.5	0.5		0.4	—	0.8	0.7	0.4	0.5
Sense of Belonging	0.6	0.5	0.3	0.6	0.4	0.3	0.4	0.5	0.6	0.5	0.5	—	0.3	0.3	0.8
Self-Management	0.5	0.5	0.6	0.4	0.6	0.3	0.5	0.5		0.4	0.6	0.4	—	0.4	0.4
School Safety	0.4	0.1	0.2	0.1	0.2	0.0	0.0	0.1	0.1	0.2	0.1	0.2	0.2	—	0.3
Teacher-Student Relationships	0.7	0.4	0.1	0.7	0.3	0.3	0.4	0.5	0.8	0.8	0.4	0.6	0.4	0.2	—

Note: We calculated Spearman rank-order correlations to minimize measurement assumptions (see Chen & Popovich, 2002). Values above the diagonal are school-level correlations (between school-level mean topic scores), and numbers below the diagonal are student-level correlations. Blank cells indicate topic pairs with insufficiently sized samples (fewer than 20 schools or 500 students).

Conclusion

Panorama developed its student surveys through a rigorous process that builds in validity at the outset. This report provides psychometric evidence for the reliability and validity of our student topics. Since nearly all students complete all items from the topics administered to them, missing responses do not undermine the quality of the data. Additionally, the topics demonstrate strong internal consistency and structural validity. Finally, intercorrelations among the topics denote discriminant and convergent validity. Panorama's surveys therefore offer schools and districts a research-backed means for measuring the life skills and school climate variables that undergird child development and academic growth.

2024 Addendum for Versioned Content

Introduction

Panorama Education's student surveys cover a range of topics about students' skills, competencies, and experiences, as well as supports & environments. Over the years, we have shown a substantial body of evidence supporting the reliability and validity of our scales.

Concurrently, we gathered feedback on our survey content from educational partners and experts, continued to review the developing literature in the field of surveying, and internally audited our content. The combination of these inputs led to our undertaking a multi-year process to update our survey content to better align with the reading and comprehension needs of students.

The results of our content versioning process demonstrate considerable evidence for the reliability and validity of our updated topics and their being better aligned with the reading and comprehension needs of students. Below, we provide highlights of the content versioning development process and the outcomes of that process.

Versioning Development Process

In the first stages of our versioning process we gathered, reviewed, and summarized the years of feedback we received on our student survey content. The feedback came from numerous constituencies, including users, practitioners, community members, and internal and external experts. A central theme that emerged from this feedback was that some of our content was challenging for students, especially younger students, to read and respond to. Our research team conducted an internal review of our student survey content, and also concluded that there were

meaningful opportunities to enhance readability[4]. Subsequently, we decided to update our student survey content, with the core goal of improving content readability for students in grades 3-5.

To support this update, we conducted a literature review on best practices for surveying young students. We used the findings of that review, combined with the historical feedback we received on our student survey content, our empirical analyses of years of survey data (representing millions of students in thousands of schools), and our expertise in survey design, to do an item-by-item audit of our 3-5 student survey content to flag items that were good candidates for revision. We also conducted a series of interviews with practitioners (elementary school teachers, reading specialists, and literacy coaches), during which they reviewed our 3-5 student content and flagged any items they saw as challenging for our youngest survey takers (3rd grade students). During these interviews, practitioners also proposed changes to the content that they felt could improve its readability. Additionally, we conducted cognitive interviews[5] with students to help us flag difficult items.

Based on those inputs, Panorama's research team drafted an initial series of revisions to the 3-5 student survey content. We then conducted another round of interviews with practitioners and experts during which they reviewed items that were previously flagged for readability issues, suggested improvements to flagged items, and evaluated the revisions to flagged items drafted by Panorama's research team. We used the feedback from those interviews to further revise items.

After determining the items that would ultimately be considered for updates in this round of versioning, we then conducted a series of cognitive interviews with elementary school students to assess the updated content. Based on the results of those interviews, we made an additional set of revisions to the proposed content and piloted our topics with districts and schools who opted-in to our content pilot program[6].

A brief description of the pilot program and the results from the program are shared below.

---

[4] In our review of student survey content, we also uncovered a small number of additional versioning opportunities, beyond readability. We decided to pursue a selected group of these opportunities. For example, to make our language more inclusive, we changed the language in an item from "if you walked into class upset…" to "if you were upset when you came into class…". Other updates pertained to improving relevancy and construct alignment. All revisions were part of the iterative review and piloting processes outlined in this section.

[5] In our cognitive interviews, students read each survey item out loud to a member of the research team and talked through their thought processes as they generated a response. The research team member sometimes asked follow-up questions. This process enabled us to gather important data on the extent to which students could read, understand, and respond to each item in the way that our research team intended.

[6] During our development process, we received considerable feedback from educational partners and practitioners about the wide variability in reading levels among older students (those in grades 6-12). Based on this feedback, along with our own analyses, we chose to revise and pilot our topics in a way that allows us to offer the same updated content across grades 3-12

Pilot Program

Panorama's survey content pilot program launched in fall 2023 to support Panorama's versioning of student survey content. Districts and schools that met predetermined criteria were given the opportunity to opt into the pilot program prior to launching their student surveys in 2023-2024. Those who opted in were assigned an updated (revised) student survey topic (that was placed at the end of their student survey), which did not overlap with the other topics in their survey. Eleven revised topics were piloted during Academic Year (AY) 2023-2024.

Results

Sample Demographics

Table 1 presents the demographics for the pilot program student sample[7]. The sample of 146,962 students shows considerable demographic diversity. As expected, it contains an approximately even distribution of students with respect to gender and grade level.

Table 1. Student Demographics	
Demographic Group	**%**
Female	49
Male	49
Asian	5
Black	13
Latinx	20
Multiracial	5
Native American	1
Pacific Islander	1
White	55
3rd Grade	9
4th Grade	10
5th Grade	11
6th Grade	10
7th Grade	10
8th Grade	10
9th Grade	11
10th Grade	11
11th Grade	10

---

[7] Since some students were missing demographic data—due to schools or districts electing to not include demographic data with their surveys or certain demographics not being recorded in school and district data systems—percentages reported reflect just those students with non-missing data for the given area. The percentage of students with missing demographics ranged from a low of 2% for student gender to a high of 43% for free/reduced-price lunch eligibility. Note that the percentages may not total to 100 due to rounding error.

12th Grade	8
English language learner	11
Eligible for free/reduced-price lunch	27
Special education	15

Reliability

Reliability, as assessed through Cronbach's alpha, is essentially a measure of signal-to-noise (DeVellis, 2016) with higher values reflecting more "signal" and less "noise." Put differently, Cronbach's alpha measures how similarly students respond to items from the same topic, i.e., a topic's internal consistency (Streiner, 2003), with higher values reflecting greater reliability. For Cronbach's alpha, the What Works Clearinghouse (2022) indicates a minimum alpha level of 0.60, though many experts report that a value of .70 or higher indicates sufficient reliability. As shown in Table 2, all versioned topics exhibited alphas of .70 or higher, with the majority above .80.

Validity

We conducted confirmatory factor analyses to assess the structural validity of each survey topic (see Messick, 1995b). More specifically, we examined whether each topic had the statistical structure we expected, i.e., whether it measures only a single dimension (and not multiple dimensions). Table 2 presents the two key statistics from each analysis: the comparative fit index (CFI) and the root mean square error of approximation (RMSEA).

With a maximum possible value of 1, CFI measures how well the data from each topic fit a one-dimensional solution; historically, a CFI of 0.90 or greater has been considered sufficient, though a value closer to 0.95 or above is preferred. RMSEA is a complementary measure of model fit, with lower values indicating better fit and a typical threshold of 0.08. (For a discussion of these metrics and their conventions, see Hu & Bentler, 1999; Laverdière et al., 2013). As shown in Table 2, all topics were above the preferred CFI threshold of 0.95, with the exception of Social Awareness, which was still above the 0.90 threshold. Social Awareness was also the only topic that was at, and not under, the typical RMSEA threshold of 0.08.

Table 2. Reliability and Structural Validity Results by Topic			
Topic	**α**	**CFI**	**RMSEA**
Climate	0.81	0.99	0.05
Emotion Regulation	0.86	0.99	0.06
Engagement	0.86	0.99	0.06
Grit (Perseverance)	0.74	0.99	0.04
Growth Mindset	0.70	0.99	0.06
Rigorous Expectations	0.79	0.99	0.05
Self-Efficacy	0.86	0.99	0.07
Self-Management	0.83	0.98	0.07
Sense of Belonging	0.80	0.99	0.06

Social Awareness	0.81	0.93	0.08
Teacher-Student Relationships	0.82	0.99	0.06

Note: α = Cronbach's alpha, CFI = comparative fit index, RMSEA = root mean square error of approximation

To address convergent and discriminant validity, we computed student-level intercorrelations among topics and compared those with what one would expect based on extant scholarship. The survey topics correlate with each other largely as expected. With only one exception (a trivially small negative correlation between Growth Mindset and School Safety), the correlations are all positive, with stronger correlations for more related constructs (e.g., Sense of Belonging and Teacher-Student Relationships) and weaker correlations for less related constructs (e.g., School Safety and Growth Mindset).

As one of many possible examples of convergent validity, and in line with the outcomes of prior analyses of our topic correlations, we see strong correlations (0.5 – 0.7) between our Teacher-Student Relationships, Rigorous Expectations, Sense of Belonging, and Climate topics. Based on prior work showing the centrality of students' relationships with their teachers to their social experience of school more generally, as well as the impact of teachers' setting high expectations for their students, we would expect these topics to be highly related (see Anderman, 2003; Brinkworth, McIntyre, Juraschek, & Gehlbach, 2018; Lee, 2012).

As an example of discriminant validity, and in line with the outcomes of prior analyses of our topic correlations, we see relatively no meaningful correlations between students' perception of the safety of their environment as measured by our School Safety topic and their life skills competencies [e.g., Social Awareness (0.1) and Growth Mindset (-0.1)], perceptions of their teachers [e.g., Rigorous Expectations (0.1)], or even feeling that they belong at school [e.g., Sense of Belonging (0.2)]. Based on prior psychometric evidence and research (e.g., Skiba et al., 2004; Thapa et al., 2013), one would expect students' safety experience to diverge from these other variables.

As an additional check, where data were available (i.e., for eight of the eleven topics), we also examined student-level correlations between the updated and historical versions of our topics using pilot students' most recent topic scores for historical versions of the scales. Correlations between updated and historical versions of the scales were extremely high (0.8-0.9), with the exception of the growth mindset scale, which showed a moderate correlation (0.4) (this was expected given the nature and scope of the changes for the growth mindset scale).

Conclusions

Panorama developed and versioned its student surveys through a rigorous process that builds in validity at the outset. This addendum provides an overview of the rigorous development process

used to improve content readability and psychometric evidence for the reliability and validity of our versioned student topics. Since nearly all students completed all items from the topics administered to them (with an item skip rate of 1% or less across all items), missing responses do not undermine the quality of the data. The versioned topics demonstrate strong internal consistency and structural validity. The intercorrelations among the topics also denote discriminant and convergent validity in-line with expectations and prior analyses. Additionally, correlations between updated and historical versions of topics were high where expected.

Panorama's versioned topics therefore continue to offer schools and districts a research-backed means for measuring the variables that undergird child development and academic growth and success.

References

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84(3),* 261.

Anderman, L. H. (2003). Academic and social perceptions as predictors of change in middle school students' sense of school belonging. *The Journal of Experimental Education*, *72*(1), 5-22.

Artino Jr., A. R., La Rochelle, J. S., Dezee, K. J., & Gehlbach, H. (2014). Developing questionnaires for educational research: AMEE Guide No. 87. *Medical Teacher*, *36*(6), 463-474.

Barge, S. & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. Research in Higher Education, 53(2), 182-200. doi:10.1007/s11162-011-9251-2.

Brinkworth, M. E., McIntyre, J., Juraschek, A. D., & Gehlbach, H. (2018). Teacher-student relationships: The positives and negatives of assessing both perspectives. *Journal of Applied Developmental Psychology*, *55*, 24-38.

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. *MET Project Research Paper*.

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412983808

Cohen, G. L., & Garcia, J. (2008). Identity, belonging, and achievement: A model, interventions, implications. *Current Directions in Psychological Science, 17(6),* 365-369.

Cornell, D. G., & Mayer, M. J. (2010). Why do school order and safety matter? *Educational Researcher, 39(1),* 7-15.

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology, 51(1),* 171-200.

DeVellis, R. F. (2016). *Scale development: Theory and applications (Vol. 26).* Sage Publications.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2014). Self-control in school-age children. *Educational Psychologist, 49(3),* 199-217.

Duckworth, A., & Gross, J. J. (2014). Self-control and grit: Related but separable determinants of success. *Current Directions in Psychological Science, 23(5),* 319-325.

Dweck, C. S. (2008). *Mindset: The new psychology of success*. Random House Digital, Inc.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28(4)*, 717-741.

Fowler Jr., F. J. (2013). *Survey research methods*. Sage publications.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74(1),* 59-109.

Gehlbach, H., & Artino, A. R. (2018). The survey checklist (manifesto). Academic Medicine: Journal of The Association of American Medical Colleges, 93(3), 360-366. doi:10.1097/ACM.0000000000002083

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*(4), 380.

Good, T. L. (2014). What do we know about how teachers influence student performance on standardized tests: And why do we know so little about other student outcomes. *Teachers College Record, 116(1),* 1-41.

Holt, D., & Elliot, D. (1991). Methods of weighting for unit non-response. *Journal of the Royal Statistical Society: Series D (The Statistician), 40*(3), 333-342.

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

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9(2),* 131-155.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5(3),* 213–236.

Krosnick, J. A. (1999a). Maximizing questionnaire quality. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Political Attitudes* (pp. 37-57). San Diego: Academic Press.

Krosnick, J. A. (1999b). Survey research. *Annual Review of Psychology*, *50*(1), 537-567.

Laverdière, O., Morin, A. J., & St-Hilaire, F. (2013). Factor structure and measurement invariance of a short measure of the Big Five personality traits. *Personality and Individual Differences*, *55*(7), 739-743.

Lee, J. S. (2012). The effects of the teacher–student relationship and academic press on student engagement and academic performance. *International Journal of Educational Research*, *53*, 330-340.

Messano, F., Childress, S., Atwood, J., & Kubagawa, M. (2019). Insight brief: Using expanded measures of student success for school improvement. Retrieved from http://www.newschools.org/wp-content/uploads/2019/10/EDSS_2019.pdf.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice*, *14*(4), 5-8.

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

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., and Barmer, A. (2019). The Condition of Education 2019 (NCES 2019-144). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144.

Osterman, K. F. (2000). Students' need for belonging in the school community. *Review of Educational Research, 70(3),* 323-367.

Pianta, R. C., Hamre, B. K., & Allen, J. P. (2012). Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In *Handbook of Research on Student Engagement* (pp. 365-386). Springer, Boston, MA.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82(1),* 33-40.

Richeson, J. A., & Sommers, S. R. (2016). Toward a social psychology of race and race relations for the twenty-first century. *Annual Review of Psychology, 67,* 439-463.

Roorda, D. L., Koomen, H. M., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81(4),* 493-529.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55(1),* 68.

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*(1), 61-79.

Skiba, R., Simmons, A. B., Peterson, R., McKelvey, J., Forde, S., & Gallini, S. (2004). Beyond guns, drugs and gangs: The structure of student perceptions of school safety. *Journal of School Violence, 3(2-3),* 149-171.

Simon, H. A. (1957). *Models of Man: Social and Rational.* New York: Wiley

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103.

Sue, D. W. (2013). Race talk: The psychology of racial dialogues. *American Psychologist, 68(8),* 663.

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

Transforming Education (2016). Measuring MESH: Student and Teacher Surveys Curated for the CORE Districts.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research, 78(4),* 751-796.

Vriesema, C. C., & Gehlbach, H. (2019). Assessing Survey Satisficing: The Impact of Unmotivated Questionnaire Respondents on Data Quality. Working Paper. Policy Analysis for California Education, PACE. Retrieved from https://edpolicyinca.org/publications/assessing-survey-satisficing

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972. doi: 10.1177/0013164404268674

What Works Clearinghouse: Procedures and Standards Handbook, Version 5.0. WWC 2022008. (2022). *What Works Clearinghouse*.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Man Yang, S., Carvalho, C. M., Hahn, P. R., ... Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573(7774),* 364-369.