Panorama Student Survey: Development and Psychometric Properties

This research report describes the core attributes of the *Panorama Student Survey*, the rigorous process through which the scales within the survey were developed, and the resultant psychometric information from the first two major pilot administrations as well as some smaller studies we have conducted on particular scales. The sections of this report demonstrate that each scale included in the *Panorama Student Survey* captures substantial variability, exceeds agreed upon standards of reliability, and demonstrates strong evidence of multiple types of validity.

This research report is written with two purposes in mind. First, the report is structured to provide guidance about what criteria are important in evaluating survey quality (see also the references for more extensive reading on key topics). Second, this report details key characteristics of the survey, its development process, and results from two initial pilot administrations so that readers can evaluate these characteristics objectively. We view the report as a living document and will continue adding to it as further studies are conducted. Feel free to check back in with us for updated versions at research@panoramaed.com.

Background

Gaining insights into classroom settings and facilitating improvements in teaching practices are particularly challenging in the current educational context. On the one hand, schools are relying on student perception surveys to make increasingly important decisions (e.g., teacher evaluation). On the other hand, the quality of measures to assess classrooms and teaching varies widely. In response, we developed the Panorama Student Survey as the first major survey instrument with the following critical properties:

- Explicitly designed from the outset to meet two goals:
  - Provide teachers with feedback that will be useful for improving practice,
  - Enable educators to monitor student attitudes, beliefs, and values that are predictive of important outcomes;
- Each scale developed through a theoretically-grounded, empirically-based design process (Gehlbach & Brinkworth, 2011) that meets or exceeds standards of academic scholarship;
- Adherence to best practices in survey design (e.g., wording items as questions rather than statements, employing response options that are directly linked to the underlying concept in each question to improve cognition, avoiding agree-disagree items, etc.);
- Customizable by schools and districts to allow for tailoring of the survey to specific school needs and teaching frameworks while retaining validity and reliability; and
- Freely available to any educator interested in improving pedagogical practice and student outcomes.

The Panorama Student Survey was developed by a team of researchers at the Harvard Graduate School of Education under the direction of Dr. Hunter Gehlbach.

Core attributes

Content

We chose the content for the *Panorama Student Survey* explicitly to address the multifaceted needs of teachers, schools, and districts. Specifically, teachers need feedback on their areas of strength (so that these might be further leveraged in the classroom) and areas targeted for improvement (so that they can take ownership over their professional development needs). Schools need information to understand which sub-groups of students face multiple risk factors and generate ideas for how to intervene. Districts increasingly seek data to facilitate comparisons between schools within the district and to make resource allocation decisions. The *Panorama Student Survey* was developed explicitly and deliberately with these uses in mind.

The current version of the *Panorama Student Survey* consists of 10 scales that educational organizations can use to meet their needs for getting feedback on students, teachers, and schools. Organizations may choose any or all 10 scales depending on their needs. These scales include students' perceptions of:

- Classroom climate – the overall feel of a class including aspects of the physical, social and psychological environment;
- Engagement – their own behavioral, cognitive, and affective investment in the subject and classroom;
- Grit – their ability to persevere through setbacks to achieve important long-term goals;
- Learning strategies – the extent to which they use metacognition and employ strategic tools to be active participants in their own learning process;
- Mindset – the extent to which they believe that they have the potential to change those factors that are central to their performance in a specific class;
- Pedagogical effectiveness – the quality and quantity of their learning from a particular teacher about that teacher's subject area;
- Rigorous expectations – whether they are being challenged by their teachers with high expectations for effort, understanding, persistence, and performance in the class;
- School belonging – the extent to which they feel that they are valued members of their school community;
- Teacher student relationships – the overall social and academic relationship between students and their teachers (Note: scales are available in a student version and a teacher version); and
- Valuing of the subject matter – how interesting, important, and useful a particular school subject seems.

Flexibility and Individualization

Because different educational organizations have different interests, goals, and priorities, the Panorama Student Survey offers flexibility to customize survey content tailored to their individual needs by selecting the scales that are most important to the organization. For example, a district that faces challenges with absences and truancy might prioritize the *school belonging* and *teacher-student relationships* scales. Meanwhile, a district that is eager to improve teaching practices might focus on *rigorous expectations* and *pedagogical effectiveness*. Many users prefer to administer the entire survey because of the important information derived from each scale.

What to look for

When selecting a survey or comparing between different survey options, it is particularly important to have a clear sense of the purpose(s) for which the survey will be used:

- Is the goal to obtain feedback regarding baseline levels of student engagement or teaching quality?
- Do you want to understand variation in classroom climate across different subject areas?
- Do you need to see how the use of learning strategies develops over time from elementary school to middle school?
- Are you trying to evaluate the effectiveness of an intervention that is supposed to endow students with a growth mindset?

With clarity of purpose, you can select the survey or scales that best fits your needs. Because different constituents within school communities frequently have different goals in mind, it is important to come to consensus with what the collective priorities are for a given data collection.

Development Process and Validity

The *Panorama Student Survey* was developed through the six-step design process developed by Gehlbach and Brinkworth (2011) (see also Artino, La Rochelle, DeZee, & Gehlbach, 2014). 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, 1995) – into each survey scale from the outset of the design process. The six steps in the process are summarized below:

- Literature review: Through an exhaustive review of the academic literature, the research team that developed the *Panorama Student Survey* identified and synthesized definitions for each of the 10 core constructs, reviewed available instruments to see how existing measures operationalized constructs, and listed indicators of each construct.
- Interviews and focus groups: The research team then conducted a series of interviews and focus groups with students to understand, from students' point of view, what the salient indicators of these constructs were and what language students used to describe them.
- Synthesis: The lists of indicators from the academic literature and from the student input were synthesized.
- Item creation: A series of items were developed to create a set of items that would measure each construct. The items were worded so as to adhere to the scientific best practices in designing survey items. The research team came to consensus on all the items for each scale.
- Expert review: Approximately 20 experts (in that particular content area) per scale completed a questionnaire about each of the proposed items. They assessed the construct relevance of each item, identified any important indicators that appeared to be absent from the items, and commented on any items that might cause challenges for subpopulations of students (especially students at lower reading levels).
- Cognitive pre-testing/interviewing: After revising the items based on the expert feedback, the research team conducted interviews with students in grades 4-12 (ages 10 – 19), from over a dozen different countries of origin, a range of socio-economic statuses, and a range of English fluency. In each cognitive interview students repeated each survey item in his or her own words, and thought out-loud as he or she generated a response. This process provides important data on the extent to which students understand each item in the way that that our research team intended.

Upon completion of these six steps and a round of revisions to the items, the scales were subjected to large-scale pilot tests (as described in detail below).

The second important attribute of each scale emerges directly from the aforementioned *item creation* step. The design of each item adheres to the science of best survey design

practices (Artino & Gehlbach, 2012; Artino, Gehlbach, & Durning, 2011; Dillman, Smyth, & Christian, 2014; Fowler, 2009; Krosnick & Presser, 2010). A host of empirically proven better and worse ways to design items exist that often go ignored in numerous surveys. For example, designing survey items as statements – particularly ones that require respondents to agree or disagree with the content of the statement – are particularly likely to inject additional measurement error into responses. Instead asking questions with response options that are linked to the underlying concept is the preferred practice (Dillman et al., 2014; Krosnick, 1999b; Saris, Revilla, Krosnick, & Shaeffer, 2010). Failing to label all response options, using numeric rather than verbal labels, using too few response options, explicitly or implicitly asking a multi-barreled item exemplify other commonly violated best practices (Artino et al., 2014; Dillman et al., 2014; Krosnick, 1999a; Weng, 2004). As a survey scale violates increasing numbers of these best practices, the amount of measurement error grows.

What to look for

When researching how a survey or set of scales was designed, key questions include:

- Did the scale developers build validity into the survey from the outset of the scale development process? Far too often, developers rely on a process (Clark & Watson, 1995; Comrey, 1988) that attends to validity only when they prune items during large-scale pilot testing. Focusing on validity only after items have been developed can lead to construct under-representation or misrepresentation (Messick, 1995).
- Did the development process rely on academic experts, prospective respondents, or both? Input in the design process from only one source will erode the content validity of the measure.
- Do the measures adhere to the scientific best practices in designing survey items? Most survey scales fail to apply many of the most critical rules of how to word items and response options. Each violation of these best practices injects measurement error into your respondents' data and compromise data quality. See the *Survey Design Checklist* in Appendix 1 for a list of some of the most important best practices.

Statistical Properties and Evidence of Validity

Before describing the psychometric properties of each scale it is important to be transparent regarding our view of evidence of validity. We view "validation" of a survey scale as an ongoing process (Messick, 1995). In other words, there is no such thing as a "validated" survey despite many survey developers making that claim about their scales or survey. Rather, over the course of multiple studies, more and more data are accumulated that give potential users of a survey increasing amounts of faith that the survey scales measure what they purport to measure, may be used for specific purposes, in specific contexts, for specific populations, etc.

Pilot samples

Our main samples are from distinct schools and school districts the southeastern United States (Sample 1) and from a large diverse high school in the southwestern United States (Sample 2). Overall, the samples include substantial representation across multiple grade levels and racial groups. The sample also includes significant populations of English language learners as well as native English speakers. Please see Table 1 below.

Table 1: Percentages of participants for each sample

	Sample 1 (N = 4225)	*Sample 2* (N = 2994)
Female	49.55	51.08
Race/Ethnicity		
American Indian	4.12	2.44
Asian	4.17	13.89
Black	17.30	8.08
Hispanic	19.12	16.73
White	39.74	49.37
Middle Eastern	1.16	--
Other	6.70	9.49
Home Language		
English	77.25	82.96
Spanish	17.26	8.17
Other	5.49	8.87

Important data points from additional samples are regularly being added to this document.

Estimates of reliability

A pre-requisite of validity is that the measure has adequate reliability. Reliability as assessed through coefficient alpha is essentially a measure of signal-to-noise (DeVellis, 2003). As shown in the section for each scale, the estimates for coefficient alpha for every scale is .70 or greater.

Factor structure

To address structural validity (Messick, 1995), we show evidence of model fit through results from confirmatory factor analysis results (specifically comparative fit indices and root mean square error of approximation). The choice of confirmatory factor analysis to determine whether a given scale measures a single construct (as opposed to measuring parts of multiple constructs) is important because this technique allows for formal testing of the hypothesis that a single factor is being measured. Thus, it is a more rigorous assessment of whether or not each scale is measuring a single underlying factor (as opposed to measuring more than one factor) than exploratory factor analysis or principal components analysis. See the section describing each scale for these results as well.

Convergent and discriminant validity

In the sections on each specific scale we report a number of correlations and statistical tests that provide additional evidence of validity for each scale. In each of the main pilot samples, students were randomly assigned to take one form of the survey or the other. In the first pilot, we randomly assigned students to take Form A or Form B so as to assess how well specific items or different wordings of the same item functioned with the remaining items on the scale. In the second pilot, students were again randomly assigned to one of two survey forms. However, this time for each form students took several scales from the *Panorama Student Survey* and several comparison scales (e.g., Dweck's mindset scale, measures from the MET study, etc.) that addressed identical or similar constructs. Each section reports evidence of convergent and discriminant validity.

What to look for:

- Typically, a ratio of .70 or greater is considered adequate reliability for a survey scale (DeVellis, 2003).
- Look for surveys using scales that have undergone confirmatory factor analysis – a more rigorous way to analyze factor structure than exploratory factor analysis or principal components analysis (Fabrigar, Wegener, MacCallum, & Strahan, 1999).
- Assessments of convergent and discriminant validity rely on a well-founded a priori predictions about what scales should correlate with a target measure more highly than others.

Validity Evidence for Classroom Climate

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our classroom climate scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

			Sample 1		Sample 2		Polychoric *r* correlations				
	Min	*Max*	*M*	*SD*	*M*	*SD*	1	2	3	4	5
1) On most days, how pleasant or unpleasant is your teacher's mood?	1	7	5.2	1.79	6.0	1.50	--	.51	.43	.60	.23
2) How fair or unfair are the rules for students in this class?	1	7	5.4	1.82	6.1	1.42	0.63	--	.41	.44	.28
3) How pleasant or unpleasant is the physical space in this classroom?	1	7	5.1	1.71	5.6	1.54	0.54	0.54	--	.43	.22
4) How positive or negative is the energy of this class?	1	7	5.0	1.74	5.8	1.47	0.57	0.55	0.50	--	.17
5) In this class, how much does the behavior of other students hurt or help your learning?	1	5	3.5	1.17	4.1	0.93	0.43	0.41	0.39	0.43	--
Overall composite			4.9	1.24	5.5	0.95					

Notes:

Sample 1 correlations are reported below the diagonal; Sample 2 correlations are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit χ^2 (5 *df*)	9.44	13.19
p	0.09	.022
RMSEA estimate (90% CI)	.023 (0, 0.045)	.032 (.011, .054)
CFI	.999	.997
Coefficient α	.789	.728
Factor loadings		
Item 1	0.765	0.708
Item 2	0.753	0.588
Item 3	0.697	0.545
Item 4	0.708	0.710
Item 5	0.567	0.275

Evidence of validity

To provide a comparison and evidence of convergent validity, in the second sample we also included a previously published, well regarded classroom climate scale. We included 18 items with a reliability estimate of $\alpha = .80$ for this sample. Our revised scale and the original scale correlated moderately ($r = .52$).

Like the comparison scale, our scale shows weak correlations with characteristics of students such as parental education, course grades, and discipline. This suggests that how students' perceive the classroom climate is not a function of aspects of their background, their academic performance, or how frequently they get into trouble at school. Thus, we find strong evidence for discriminant validity between our *Classroom Climate* scale and these other variables. By contrast, our scale has significantly stronger correlations than our comparison scale with two constructs that are theoretically related: *Rigorous expectations* and *Pedagogical Effectiveness*. This evidence suggests that our scale functions especially well for capturing elements of classroom climate related to other things that promote learning, which is our focus, and, in fact do so significantly better than other well-reputed scales.

	Anticipated correlation	*r* with our climate scale	*r* with the alternate climate scale	Significant difference between correlations
Mother's education level	none	-.005	.008	$p = .80$
Course grades	weak	.032	.046	$p = .76$
Discipline	weak	-.007	.023	$p = .48$
Teacher press	moderate	.529	.312	$p < 0.001$
Pedagogical Effectiveness	moderate	.583	.336	$p < 0.001$

Within the first sample, *Classroom Climate* also correlated strongly with *Teacher-student relationships* $r = .64$ [95% CI: .61, .67], and *Engagement* $r = .63$ [95% CI: .60, .66]. This is again consistent with our general goal of measuring various aspects of students' classroom experiences around learning.

Validity Evidence for Engagement

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our revised engagement scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric correlations				
	M	*SD*	*M*	*SD*	1	2	3	4	5
1) In this class, how eager are you to participate?	3.9	1.0	3.3	0.9	--	0.27	0.32	0.36	0.36
2) When you are not in class, how often do you talk about ideas from class?	2.7	1.2	2.2	1.0	0.29	--	0.42	0.47	0.52
3) How often do you get so focused on class activities that you lose track of time?	3.2	1.2	2.8	1.1	0.31	0.46	--	0.48	0.54
4) How excited are you about going to this class?	3.4	1.3	2.7	1.2	0.43	0.52	0.45	--	0.76
5) Overall, how interested are you in this class?	3.5	1.2	3.0	1.1	0.39	0.57	0.42	0.71	--
Overall composite	3.4	0.9	2.7	0.8					

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit χ^2 (5 *df*)	86.901	24.977
p	<.001	.001
RMSEA estimate (90% CI)	.095 (.078, .112)	.05 (.032, .071)
CFI	.984	.996
Coefficient α (90% CI)	.78 (.76, .80)	.78 (.76, .78)
Factor loadings		
Item 1	.484	.426
Item 2	.680	.607
Item 3	.570	.621
Item 4	.831	.833
Item 5	.832	.890

Evidence of validity

In the second sample, we correlated our scale against other measures of theoretical interest. If our scale is measuring the latent construct of engagement with fidelity, we would expect our measure to correlate more highly with a series of other outcomes *that engagement is theoretically expected to predict*. Based on prior research, we know that engagement is associated with age. In Sample 2, we observed engagement falling slowly with age (r = -.13). This is consistent with Klem & Connell's (2004) finding that students become less engaged as they continue their schooling. In addition, we found that boys report higher levels of engagement than girls (about 2 tenths of a standard deviation).

We have also observed within this sample that students who indicate they are engaged in a class also tend to report that they value that particular class (r = .64). Furthermore, students who report higher engagement in class also tend to report higher levels of TSR (r = .52). There is a small to moderate correlation with grit (r = .30), and a weak association with the previous year's math grades (r = .10). This last correlation indicates that a student's engagement in a class this year does not necessarily mean that the student was engaged in the subject in the last school year. Finally, as one would expect, students who are more engaged are tardy less often (r = -.11).

Grit

The concept of grit, popularized by the research of Dr. Angela Duckworth, emerges from the notion that students who persevere towards an important, long-term goal that they are passionate about are particularly likely to succeed across many domains in life (Duckworth & Gross, 2014). In consultation with Dr. Duckworth, we developed a revision to the original grit scale (Duckworth & Quinn, 2009) that conceptualized grit very similarly as: Students' ability to persevere through setbacks to achieve important long-term goals. However, by leveraging the previously described survey design process, we aimed to reduce the measurement error and create a shorter scale (our revised scale is 6 rather than 8 items).

Table 1: Basic descriptive statistics for revised grit scale – Item means, standard deviations, and inter-item correlations.

	Sample 1		Sample 2		Pearson *r* correlations					
	M	*SD*	*M*	*SD*	1	2	3	4	5	6
1) If you have a problem while working towards an important goal, how well can you keep working?	3.7	0.98	3.4	0.87	--	.46	.43	.45	.52	.46
2) How often do you stay focused on the same goal for several months at a time?	3.5	1.11	3.1	1.01	.43	--	.44	.32	.44	.36
3) Some people pursue some of their goals for a long time, and others change their goals frequently. Over the next several years, how likely are you to continue to pursue one of your current goals?	3.9	1.06	3.6	1.01	.42	.41	--	.36	.50	.43
4) When you are working on a project that matters a lot to you, how focused can you stay when there are lots of distractions?	3.5	1.08	3.7	1.01	.41	.31	.31	--	.32	.34
5) If you fail to reach an important goal, how likely are you to try again?	4.0	1.00	3.6	0.92	.47	.41	.45	.36	--	.50
6) How likely is it that you can motivate yourself to do unpleasant tasks if they will help you accomplish your goals?	3.6	1.1	3.3	0.93	.41	.34	.36	.30	.44	--
Overall composite	**3.70**	**.71**	**3.5**	**.67**						

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit χ^2 (9 *df*)	25.68	54.53
p	.002	0
RMSEA estimate (90% CI)	.033 (.018, .048)	.057 (.043, .072)
CFI	.996	.987
Coefficient α (90% CI)	.75 (.73, .77)	. 78 (.76, .80)
Standardized factor loadings		
Item 1	0.695	0.715
Item 2	0.61	0.628
Item 3	0.628	0.672
Item 4	0.531	0.509
Item 5	0.695	0.713
Item 6	0.588	0.653

To provide a comparison and evidence of convergent validity, in the second sample we also included Duckworth and Quinn's (2009) grit scale. Their scale includes 8 items and has a reliability estimate of α = .66 for this sample. Our revised scale and her original scale correlated moderately (*r* = .53). After correcting for attenuation, the correlation was .74. This suggests that the two measures are largely measuring quite similar underlying constructs, but that measurement error, particularly from the original scale, is depressing the association between the two measures.

Evidence of validity

If our scale is measuring the latent construct of grit with more fidelity than the previous version of the scale, we would expect our measure to correlate more highly with a series of other outcomes *that grit is theoretically expected to predict* than the existing version of grit. Based on prior research (Duckworth, Kirby, Tsukayama, Berstein, & Ericsson, 2011; Duckworth & Quinn, 2009), we know that grit is associated with academic achievement. In our sample, we find that grittier students tend to achieve higher course grades overall. When using our revised grit scale we find comparable correlations with the original scale with one exception: our revised scale does a significantly better job of predicting the extent to which students will value a particular subject matter (math in this case) than the original grit scale.

	Anticipated correlation	*r* with our grit scale	*r* with the previous grit scale	Significant difference between correlations
Course grades	small to moderate	.14	.19	$p = .18$
Attendance	small	-.08	-.04	$p = .50$
Tardies	small	-.03	-.04	$p = .79$
Valuing of math	small	.32	.18	$p < .001$
Age	none	-.05	-.05	$p = .93$
Mother's education level	small	.10	.04	$p = .11$
Father's education level	small	.09	.05	$p = .25$

Validity Evidence for Learning Strategies

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our learning strategies scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric *r* correlations				
	M	*SD*	*M*	*SD*	1	2	3	4	5
1) When you get stuck while learning something new, how likely are you to try a different strategy?	3.5	1.10	3.2	1.0	--	.42	.43	.52	.51
2) How confident are you that you can choose an effective strategy to get your schoolwork done well?	3.8	0.99	3.4	1.0	.38	--	.51	.57	.55
3) Before you start on a challenging project, how often do you think about the best way to approach the project?	3.5	1.10	3.3	1.2	.41	.38	--	.46	.57
4) Overall, how well do your learning strategies help you learn more effectively?	3.7	0.99	3.4	0.9	.39	.46	.34	--	.64
5) How often do you use strategies to learn more effectively?	3.7	1.02	3.2	1.0	.47	.46	.38	.49	--
Overall composite	3.6	.74	3.3	.77					

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit χ^2 (5 *df*)	34.495	32.599
p	<.001	<.001
RMSEA estimate (90% CI)	.058 (.041, .077)	.059 (.041, .079)
CFI	.991	.995
Coefficient α (95% CI)	.74 (.72, .77)	.81 (.79, .84)
Factor loadings		
Item 1	0.633	0.641
Item 2	0.655	0.716
Item 3	0.570	0.675
Item 4	0.657	0.776
Item 5	0.721	0.812

The Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich & de Groot, 1990) is a commonly used survey for measuring a student's attitudes and competencies. One section of this survey (Part D - Cognitive Strategy Use) specifically relates to learning strategies. To provide a comparison and evidence of convergent validity, we also included this section of the MSLQ in the second sample. This scale includes 13 items and has a reliability estimate of $\alpha = .86$ for this sample. Considering that our scale has only 5 items, its reliability estimate of $\alpha = .81$ is remarkably high in comparison. Our revised scale and this commonly used scale have a moderately strong correlation ($r = .64$). After correcting for attenuation, the correlation was .78. This suggests that the two measures are measuring highly related constructs, but we believe that our measurement approach is capturing a valuable difference.

Evidence of validity

Students with greater learning strategies skills are expected to achieve more highly academically (Pintrich, Smith, Garcia, & McKeachie, 1993). Our new learning strategies scale is associated with the grades that students earn in a similar manner as the well established MSLQ scale (see Table 3). Similarly, we anticipate that students who better employ learning strategies would engage more in classroom activities. Our scale finds a higher correlation with engagement than the MSLQ (r = .45 vs r = .41). We also anticipate that these same students should value their class to a greater degree than those who are less adept at using learning strategies – an association that our scale shows more strongly (r = .50 vs r = .39) than the MSLQ.

Grit measures whether a student will stick with important tasks through setbacks. We expect that students with greater learning strategy skills may have more strategies at their disposal to help them to persevere. Our scale shows a moderately large association between these two constructs (r = .52). Our learning strategy scale also demonstrates a correlation with mindset (r = .37), which is a predicted association since a growth mindset means that a student believes that they have control over changing their performance. On the one hand, we might expect students to become more strategic as they age; on the other hand, students typically become less motivated over time. We find that our scale is *not* related to a student's age (r = -.01, not statistically significant).

Table 3: Comparison of Learning Strategies Scales

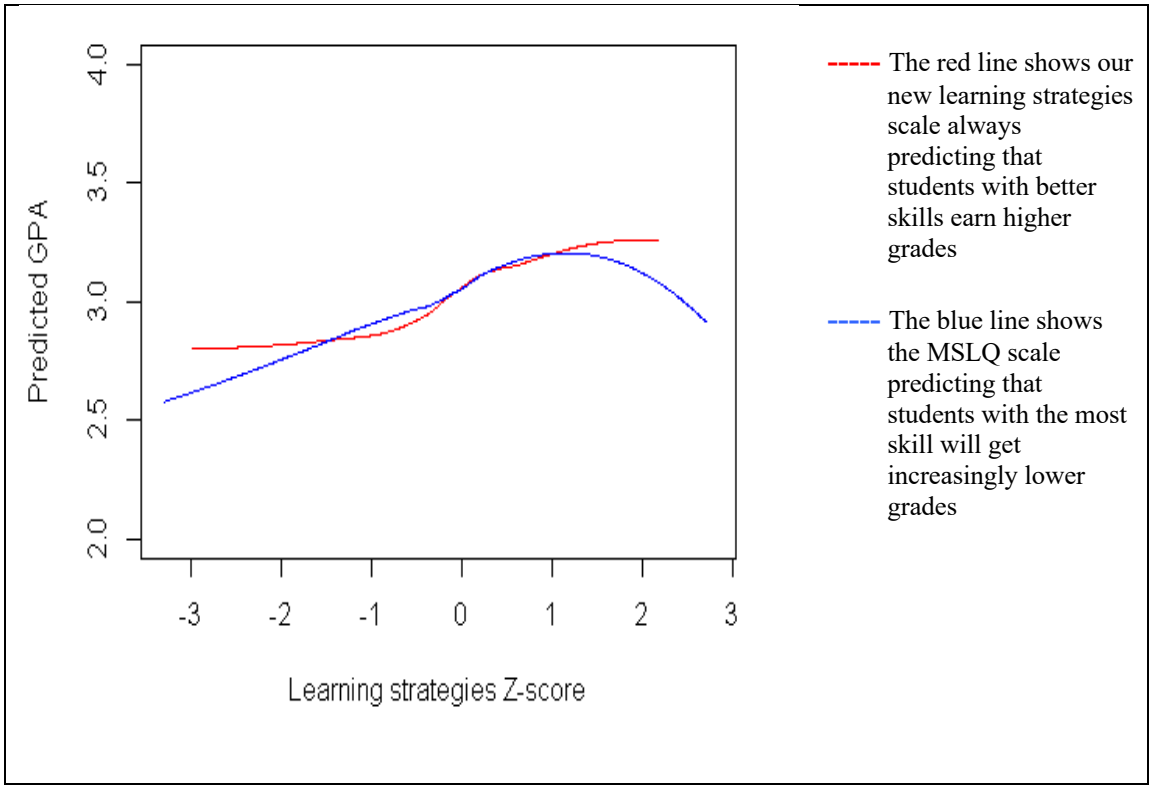
	Anticipated relationship	Correlation with our learning strategies scale	Correlation with the MSLQ learning strategies scale	Significant difference between correlations
Course Grades	small	.18*	.16*	*p*=.464
Classroom Engagement	moderate	.45*	.41*	*p=.030***
Valuing Math	moderate	.50*	.39*	*p<.001***
Grit	moderate	.52*	.48*	*p=.030***
Mindset	moderate	.37*	.32*	*p=.011***
Age	none	-.01	-.06*	*p=.028***

Notes: * indicates that the correlation is statistically significant at $p < .05$

** indicates that the difference between the two correlations is statistically significant at $p < .05$

Beyond the strength of the correlation, it is also instructive to understand the type of association the two learning strategies scales have with students' grades. The MSLQ scale items appropriately predict student grades for low achieving students, but this association becomes less clear for moderate and high achieving students. In other words, what the MSLQ measures suggests that the highest achieving students have no better (and even slightly worse) learning strategy skills (see blue line in Figure 1) – which seems unlikely. We believe that this finding may be an artifact due to the "laundry list" approach that simply measures how many strategies students claim to use. In contrast, our new scale measures how well students understand *how to use* their learning strategies – we believe that this better assesses their skill and will better relate to their achievement. Our approach means that we can continue to distinguish how learning strategy skills relates to academic achievement – for all levels of students (see red line in Figure 1).

Figure 1: Relationship between the two learning strategies scales and course grades



Validity Evidence for Mindset

It is important to bear in mind key conceptual differences in our approach to measuring mindset as compared to Dweck's (1975) original conception. Specifically, our revised mindset scale focuses broadly on those traits, dispositions, and capacities that facilitate students' achievement in a given class. By contrast, the original mindset scale that we compare our measure against in the second sample takes a narrower approach – focusing exclusively on intelligence. For both scales, the emphasis is on how changeable students perceive these capacities to be.

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our revised mindset scale. For our revised scale, results between the two samples differ regarding both the variability between each item, and the correlations between items. The different age ranges of the two samples may explain this difference. Sample 1 included younger and older students and a range of schools, whereas sample 2 consists of primarily older children from a single high school. If school climate, peers, and pedagogical approaches influence perceived malleability, it would make sense that the second sample manifests slightly less variability.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric *r* correlations					
	M	*SD*	*M*	*SD*	1	2	3	4	5	6
1) Being talented	3.0	1.3	2.7	1.13	--	.07	.19	.08	.34	.27
2) Putting forth a lot of effort	3.7	1.3	4.2	0.99	.38	--	.29	.64	.18	.48
3) Having support from other people	3.3	1.3	3.2	1.06	.41	.50	--	.25	.27	.33
4) Behaving well in class	3.7	1.4	4.0	1.12	.31	.73	.48	--	.17	.45
5) Liking the subject	3.1	1.3	2.7	1.12	.42	.38	.44	.40	--	.38
6) Being able to overcome obstacles	3.5	1.3	3.5	0.99	.41	.66	.51	.62	.52	--
Overall composite	**3.4**	**0.95**	**3.4**	**0.66**						

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and model fit

	Sample 1	Sample 2
Model fit χ^2 (9 *df*)	292.18	279.81
p	< 0.001	< 0.001
RMSEA estimate (90% CI)	.139 (.125, .153)	.138 (.124, .152)
CFI	.961	.924
Coefficient α (95% CI)	.81 (.79, .83)	.67 (.64, .69)
Standardized factor loadings		
Item 1	.523	.293
Item 2	.835	.772
Item 3	.650	.438
Item 4	.804	.746
Item 5	.600	.470
Item 6	.792	.671

Dweck's (2006) mindset scale was also included in the second sample. Her scale includes 4 items and has a reliability estimate of $\alpha = .89$ for this sample. Our revised scale and her original scale correlated modestly ($r = .29$). After correcting for attenuation, the correlation was .38. This suggests that the two scales are measuring related but different underlying constructs as explained in the introduction to this section.

Evidence of validity

Because mindset is the extent to which students believe that their performance is something that they can change (rather than a measurement of who and what they innately are), we expect this scale to correlate with certain other aspects of the classroom. In other words, we anticipate that students with higher mindset scores (more of a "growth" mindset) will engage with their learning and see challenges as learning opportunities, despite setbacks. Specifically, having a growth mindset should make students eager to:

- gain and employ learning strategies (*because students should believe that learning can improve with effort*),
- engage in classroom activities (*because students should believe that engaging in classroom discussions and activities might promote learning*),
- value the subject that they are studying (*because students should see learning as a process, and class as a valuable learning opportunity*) and
- stick with an important goal through difficulties (*because if students believed their abilities were fixed, there would be no point in persisting*).

In addition, although only small correlations are expected for course grades, age, and discipline, Table 3 shows that these associations are all in the expected direction. In other words, students with more of a growth mindset have a slight tendency to get better grades and be involved in fewer disciplinary incidents – both of which would be expected from theory and previous studies.

Table 3: Comparison of scales for Sample 2

	Anticipated correlation	*r* with our mindset scale	*r* with the previous mindset scale	Significant difference between correlations
Attendance	none	.04	-.03	$p = .16$
Course grades	small	.06*	.06*	$p = .98$
Age	small	-.03	-.06*	$p = .37$
Discipline	small	-.05*	-.02	$p = .22$
Parental education level	small	.06*	.03	$p = .30$
Learning strategies	moderate	.37*	.28*	$p = .07$
Student Engagement	moderate	.34*	.17*	$p < .01$**
Valuing of Math	moderate	.39*	.22*	$p < .01$**
Grit	moderate	.30*	.30*	$p = .74$

Notes: * indicates that the correlation is statistically significant at $p < .05$

** indicates that the difference between the two correlations is statistically significant at $p < .05$

An additional set of evidence

Thanks to a collaboration with Transforming Education, we also had an additional opportunity to compare how our revised mindset scale compared to the original Dweck items in a study they conducted. They conducted a large survey administration in grades 5-12 in 18 schools from 6 urban districts in the western U.S. Students were split across two forms of the survey, with approximately half of the students taking our revised mindset scale (N = 3822) and the remainder taking Dweck's original scale (N = 3824). From this sample, we find that our revised mindset scale had a reliability of α = .78 and Dweck's had a reliability of α = .68. A comparison of the scales can be found in Table 4 below:

Table 4: Comparison of scales from Transforming Education study

	Anticipated correlation	*r* with our mindset scale	*r* with the previous mindset scale	Significant difference between correlations
Absences	none	-.07*	-.07*	p = .79
Math standardized test (middle school students)	small	.09*	.29*	p <.001**
ELA standardized test (middle school students)	small	.09*	.32*	p <.001**
Has student been suspended?	small	-.11*	-.04	p = .04*
Year-end GPA (high school students)	moderate	.16*	.20*	p = .41
Self-reported classroom effort	moderate	.27*	.10*	p <.001**

Notes: * indicates that the correlation is statistically significant at $p < .05$

** indicates that the difference between the two correlations is statistically significant at $p < .05$

These additional, independent results suggest that:

- The previous mindset scale, with its more the narrow focus on the malleability of intelligence, is a better predictor of standardized tests than our revised scale.
- Our revised scale, with its broader focus, predicts other outcomes as well or better than the original scale.

Validity Evidence for Pedagogical Effectiveness

Basic Descriptive Statistics

Table 1 displays basic descriptive statistics for our pedagogical effectiveness items. With both Sample 1 and 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items. The correlations are in expected directions given that with both Samples items 9-13 were worded negatively (although they have since been reworded positively to improve measurement properties).

Table 1. Pedagogical effectiveness scale: Item means, standard deviations, and polychoric corrrelations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric correlations															CCSR		Tripod/MET
Pedagogical Effectiveness Items	*M*	*SD*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AP	Dist.	Challenge
1 Overall, how much have you learned from this teacher about <SUBJECT>?	3.9	1.12	3.7	0.96	-	.59	.70	.61	.58	.62	.62	.51	-.10	-.28	-.19	-.27	-.21	.74	.82	.61	-.14	.53
2 During class, how motivating are the activities that this teacher has you do?	3.4	1.17	2.9	1.13	.57	-	.67	.59	.78	.59	.57	.50	.09	-.27	-.17	-.14	-.07	.64	.66	.61	-.09	.48
3 How good is this teacher at teaching in the way that you personally learn best?	3.9	1.15	3.5	1.09	.65	.62	-	.75	.67	.66	.66	.53	-.02	-.38	-.31	-.22	-.19	.80	.81	.67	-.12	.54
4 For this class, how clearly does this teacher present the information that you need to learn?	3.9	1.07	3.8	1.03	.65	.62	.72	-	.62	.68	.64	.52	-.05	-.45	-.34	-.24	-.28	.78	.76	.61	-.11	.55
5 How interesting does this teacher make what you are learning in class?	3.5	1.23	3.1	1.2	.59	.66	.69	.63	-	.61	.60	.50	.04	-.29	-.20	-.12	-.07	.65	.66	.62	-.13	.49
6 How often does this teacher give you feedback that helps you learn?	3.7	1.19	3.8	0.93	.55	.53	.58	.54	.50	-	.66	.53	-.11	-.34	-.25	-.25	-.24	.69	.68	.68	-.16	.63
7 When you need extra help, how good is this teacher at giving you that help?	4	1.13	3.9	0.97	.62	.56	.73	.73	.64	.60	-	.60	-.11	-.36	-.28	-.24	-.25	.75	.71	.72	-.15	.53
8 How comfortable are you asking this teacher questions about what you are learning in his or her class?	3.6	1.24	3.7	1.09	.47	.45	.52	.51	.44	.51	.58	-	-.02	-.36	-.22	-.18	-.19	.58	.55	.55	-.15	.43
9 During class, how good is this teacher at making sure students do **not** get out of control?	2.1	1.29	1.6	0.92	-.32	-.17	-.30	-.33	-.18	-.25	-.34	-.11	-	.26	.21	.56	.37	-.10	-.10	-.11	.57	-.24
10 In a typical class, how clearly do you understand what this teacher expects of you?	2.6	1.28	2.4	1.06	-.10	.05	-.06	-.14	-.01	-.03	-.10	-.14	.32	-	.41	.24	.41	-.41	-.40	-.34	.27	-.29
11 How well can this teacher tell whether or not you understand a topic?	2.7	1.44	2.5	1.24	-.09	.00	-.07	-.11	-.01	-.05	-.08	-.07	.37	.35	-	.24	.37	-.32	-.28	-.24	.21	-.19
12 How good is this teacher at making sure time does **not** get wasted in this class?	2.6	1.24	2	1	-.22	-.15	-.19	-.21	-.17	-.16	-.23	-.16	.52	.39	.32	-	.36	-.27	-.26	-.25	.45	-.34
13 How much does this teacher know about the topic of his or her class?	2.3	1.31	1.7	1	-.10	.04	-.03	-.09	.04	-.05	-.09	-.10	.45	.46	.43	.41	-	-.32	-.26	-.19	.28	-.26
14 How good is this teacher at helping you learn about <SUBJECT>?			3.9	1.02														-	.86	.71	-.14	.59
15 How well has this teacher taught you about <SUBJECT>?			3.8	0.98															-	.69	-.16	.59
Pre-existing Scales																						
CCSR: Academic Personalism			3.02	0.52																-	-.16	.58
CCSR: Distractions			1.87	0.6																	-	-.22

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal. Item texts represent revised wording based on the first two pilots. Most notably, items 9-13 were negatively worded with Samples 1 and 2. For all items, min=1 and max=5, except for Chicago scales, min=1 and max=4.

Convergent/Divergent Validity

If our scale is indeed measuring components of pedagogical effectiveness, we would expect responses to correlate with other pre-existing scales designed to measure elements of teaching quality. With Sample 2, in addition to administering our own scale, we administered two scales created by the University of Chicago's Consortium on Chicago School Research (CCSR) and one of the teaching-focused scales administered in the Measures of Effective Teaching (MET) study.

Overall, the scale scores for each of the three pre-existing scales correlated in expected directions with our individual pedagogical effectiveness items. For example, the CCSR Academic Personalization scale correlated highly with our item regarding the teacher's ability to teach in the way a student "personally learns best" (r = .67).

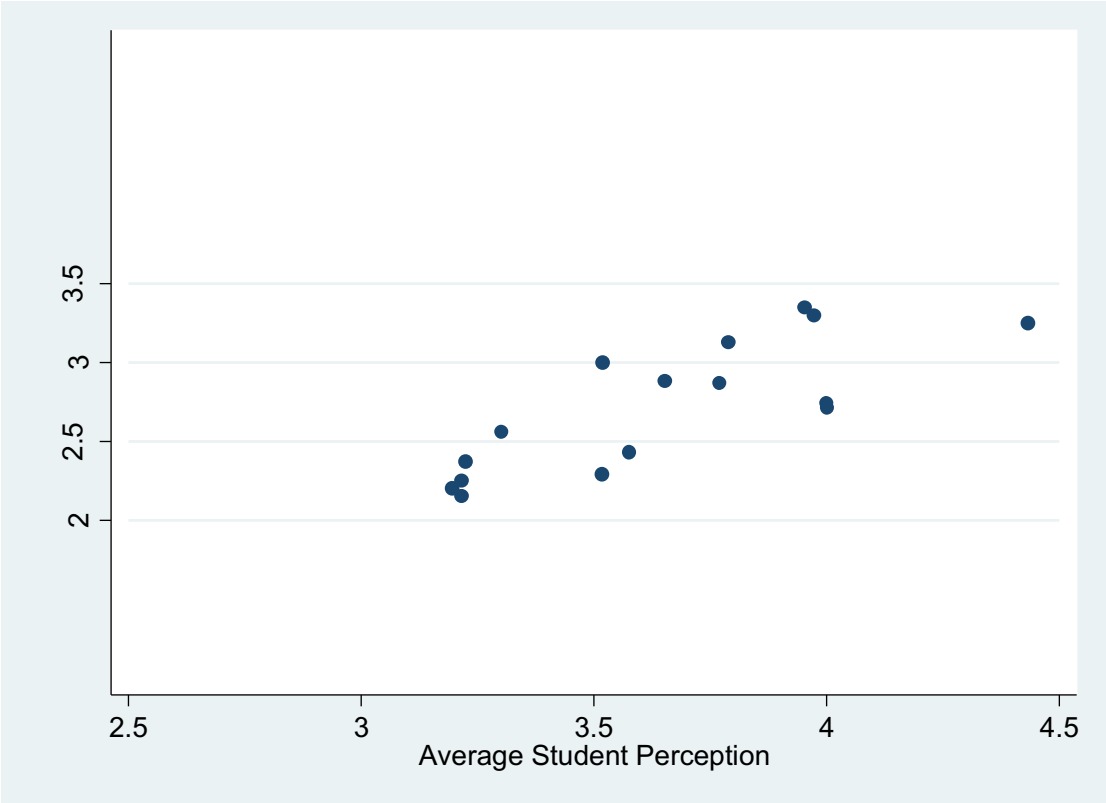
Similarly, the CCSR scale designed to measure distractions in a given classroom was most highly correlated with our item about the teacher's ability to prevent students from getting "out of control" (r = .57) and to prevent time from getting wasted (r = .45).

Finally, the MET scale designed to measure the degree to which a teacher challenges his or her students was strongly correlated with our items regarding the overall degree to which students believed they had learned from their teacher (r = .59) as well as the amount of useful feedback that teacher provides (r = .63).

Overall, these correlations are consistent with the idea that our items do indeed capture student perceptions of some of the key dimensions of teachers' pedagogical effectiveness.

Correlations with observations:

In a recent study, with a diverse (mostly Black and Latino), small, Catholic high school, we had the opportunity to correlate students' scores on the *Pedagogical Effectiveness* scale with scores from administrator observations. Students completed the survey scale for each of their (usually 5) teachers. Those scores were than averaged for each teacher across all of his or her classes so that each teacher had a single score that was represented by the aggregate ratings of nearly all of his or her students (a small percentage of students did not take the survey). Administrators performed brief, but frequent observations of their teachers over the course of the school year (typically about 10 minute observations, 10 times per year) using an adaptation of Kim Marshall's framework. Those observation ratings were then averaged so that each teacher had a single observation score. We found the correlation between the aggregated survey scores and administrators' aggregated observation scores was r = .80. The scatterplot below suggests that this high correlation was not merely due to an outlier in the small sample:


3.5
3
2.5
2
2.5
3
3.5
4
4.5
Average Student Perception

Validity Evidence for Rigorous Expectations

The extent to which teachers establish high expectations for their students and then hold students to those high expectations is pivotal for student learning. We define rigorous expectations as students' perceptions that they are being challenged by their teachers with high expectations for putting forth effort, acquiring understanding of key concepts, persisting, and performing at a high level in class.

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our rigorous expectations scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric correlations				
	M	*SD*	*M*	*SD*	1	2	3	4	5
1) How much does this teacher encourage you to do your best?	4	1.1	3.8	1.1	--	0.58	0.45	0.28	0.61
2)When you feel like giving up on a difficult task, how likely is it that this teacher will make you keep trying?	3.9	1.1	3.9	1	0.64	--	0.4	0.4	0.57
3)Overall, how high are this teacher's expectations of you?	3.9	1	3.7	0.9	0.57	0.56	--	0.36	0.39
4)How often does this teacher make you explain your answers?	3.7	1.1	4	1	0.48	0.5	0.42	--	0.3
5) How often does this teacher take time to make sure you understand the material?	3.9	1.1	4	1	0.65	0.6	0.5	0.47	--
Overall composite	3.9	0.9	3.9	0.7					

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit: $\chi^2_{df=5}$	10.7	36.9
p	0.057	<.001
RMSEA estimate (90% CI)	.026 (.000, .048)	.086 (.061, .112)
CFI	0.999	0.978
Coefficient α (95% CI)	.82 (.80,.84)	.76 (.73,.79)
Factor loadings		
Item 1	0.816	0.78
Item 2	0.793	0.75
Item 3	0.699	0.56
Item 4	0.61	0.47
Item 5	0.768	0.75

Evidence of validity

To provide a comparison and evidence of convergent validity, in the second sample we also included a comparison scale. The comparison scale had a comparable coefficient alpha (α = .79). However, the fit of this alternative scale was problematic (RMSEA = .129, CFI=.922). The correlation between the two scales was r = .75 and was r = 1.00 after correcting for attenuation.

Validity Evidence for School Belonging

A sense of belonging at school is a critical pre-requisite for most students to feel comfortable learning. We define school belonging as the extent to which students feel that they are valued members of their school community.

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our school belonging scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric correlations				
	M	*SD*	*M*	*SD*	1	2	3	4	5
1) Overall, how much do you feel like you belong at your school?	3.7	1.2	3.5	1.09	--	.68	.55	.47	.65
2) How well do people at your school understand you?	3.4	1.2	3.2	0.99	.61	--	.56	.44	.68
3) How much respect do students in your school show you?	3.5	1.2	3.4	0.94	.61	.59	--	.38	.61
4) How connected do you feel to the adults at your school?	3.3	1.2	2.9	1.05	.53	.49	.44	--	.43
5) How much do you matter to others at this school?	3.5	1.2	3.3	0.99	.58	.59	.55	.44	--
Overall composite	3.5	0.92	3.3	0.78					

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 correlations are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit: $\chi^2_{df=5}$	23.85	33.01
p	<.001	<.001
RMSEA estimate (90% CI)	.046 (.029, .065)	.059 (.041, .079)
CFI	.997	.995
Coefficient α (95% CI)	.83 (.81, .85)	. 83 (.81, .85)
Factor loadings		
Item 1	.806	.815
Item 2	.783	.824
Item 3	.746	.699
Item 4	.627	.544
Item 5	.732	.822

Evidence of validity

To provide a comparison and evidence of convergent validity, in the second sample we also included Haborg's (1998) 11-item version of the Psychological Sense of School Membership scale (Goodenow, 1993). His scale has a reliability estimate of α = .85 for this sample. Our revised scale and his scale correlated strongly (r = .81). After correcting for attenuation using a structural equations model, the correlation between the latent factors measured by the two scales was .97. This suggests that the two measures are measuring almost exactly the same underlying construct. As a result of this extremely high correlation, we report only the correlations between our scale and other measures for the remainder of this school belonging section – the correlations are essentially the same whether our scale or the Haborg scale is used.

As evidence of divergent validity, we estimated the correlations between our school belonging scale with students' reported sense of the classroom climate. We expected that school belonging would correlate moderately with classroom climate, since the former is measuring a school-level perception, and the latter is measuring a classroom level perception. The observed correlation (r = .40) is congruent with our hypothesis.

Numerous studies have demonstrated a positive association between students' school belonging and academic achievement (Osterman, 2000). Therefore, we expected our school belonging scale to predict students grades as recorded on their transcripts. We found a small, statistically significant, positive correlation between school belonging, as measured by our scale, and

students' grade point average (GPA) from the previous year ($r = 0.10$). The partial correlation between school belonging and grades increase after partialling out student gender, race, and age ($r = 0.14$). We anticipate that these associations would be stronger for current year GPA.

Given research that has shown that school belonging predicts higher levels of school participation, we also expected that students with a high school belonging might have fewer absences from school. We found a small, statistically significant association such that students who reported a greater school belonging had fewer absences ($r = -0.09$), which was similar to the result after partialling out student race, gender, and grade.

Next, we anticipated that students who perceived a greater school belonging at school would have fewer disciplinary incidents at school. For example, students who get in trouble more frequently may feel increasingly like they do not belong. We found that students who reported greater school belonging did end up being reported for significantly fewer disciplinary incidents ($r = -0.06$).

Finally, we predicted that students who enrolled in more challenging courses would tend to feel greater school belonging. Although we were not formally able to test this hypothesis because we did not have precise data to categorize certain classes as definitively more advanced than others, we did explore this proposition informally. As expected, students in more advanced mathematics classes reported higher values for school belonging than their counterparts in less advanced classes. The trends were the same for both scales.

Overall, our school belonging scale performs quite similarly to the Haborg scale which was designed to measure the same construct. Yet, our school belonging scale has only five items as compared to the 11-item Haborg scale. Thus, there is a substantial gain in efficiency through using the Panorama School Belonging scale.

Validity Evidence for Teacher-Student Relationships

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our revised TSR scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric correlations				
	M	*SD*	*M*	*SD*	1	2	3	4	5
1) When your teacher asks how you are doing, how often do you feel that your teacher is really interested in your answer?	3.6	1.2	3.2	1.2	--	0.59	0.64	0.61	0.61
2) How interested is this teacher in what you do outside of class?	3.0	1.2	2.2	1.1	0.51	--	0.64	0.61	0.61
3) If you walked into class upset, how concerned would your teacher be?	3.5	1.3	2.9	1.1	0.63	0.55	--	0.64	0.68
4) If you came back to visit class three years from now, how excited would this teacher be to see you?	3.6	1.3	2.8	1.2	0.6	0.53	0.61	--	0.68
5) If you tried to tell your teacher about something that was on your mind, how carefully would this teacher listen to you?	3.8	1.2	3.3	1.1	0.64	0.5	0.63	0.65	--
Overall composite	3.5	1.0	2.8	0.9					

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit χ^2 (5 *df*)	26.555	32.706
p	.0001	<.0001
RMSEA estimate (90% CI)	.49 (.032, .069)	.059 (.041, .079)
CFI	.997	.993
Coefficient α (90% CI)	.85 (.83, .86)	.87 (.85, .89)
Factor loadings		
Item 1	.783	.799
Item 2	.664	.764
Item 3	.795	.820
Item 4	.783	.765
Item 5	.802	.820

Evidence of validity

In sample 2 we compared our scale to an existing measure of teacher caring that was used in the MET study. The 3-item MET scale (α=.80) correlates with our scale r = .85 (r = 1.00 after correcting for attenuation). Thus, the two scales are measuring very similar constructs.

If our scale is measuring the latent construct of TSR with fidelity, we would expect our measure to correlate more highly with a series of other outcomes *that TSR is theoretically expected to predict*. Specifically, we predict that students whose teacher takes more of an interest in his/her students and fosters better relationships with them will: be more engaged, employ learning strategies more regularly, value the subject matter more, and be absent less frequently. We find these expectations are largely upheld:

1) Engagement: r = .49 (p < .001)
2) Learning strategies: r = .27 (p < .001)
3) Valuing of the subject matter: r = .35 (p < .001)
4) Absences: r = -.07 (p < .10)

Validity Evidence for Valuing of the Subject Matter

Basic Descriptive Statistics

This first section presents basic descriptive statistics for our revised valuing scale. For both Sample 1 and Sample 2 (as described in the overview) each item shows substantial variability and moderately strong correlations between each of the items.

Table 1: Item means, standard deviations, and inter-item correlations for Samples 1 & 2.

	Sample 1		Sample 2		Polychoric correlations				
	M	*SD*	*M*	*SD*	1	2	3	4	5
1) How often do you use ideas from [SUBJECT] in your daily life?	3.1	1.2	2.6	1.6	--	0.65	0.28	0.6	0.5
2) How useful do you think [SUBJECT] will be to you in the future?	3.8	1.2	3.7	1.18	0.51	--	0.44	0.55	0.54
3) How important is it to you to do well in [SUBJECT]?	4.1	1	4.3	0.82	0.46	0.48	--	0.36	0.36
4) How interesting do you find the things you learn in [SUBJECT]?	3.5	1.1	2.8	1.14	0.54	0.56	0.48	--	0.58
5) How much do you see yourself as a/an [SUBJECT] person?	3	1.3	2.9	1.25	0.48	0.48	0.38	0.57	--
Overall composite	3.5	0.9	3.2	0.8					

Notes: Sample 1 correlations are reported below the diagonal; Sample 2 means are reported above the diagonal.

Students view different subject areas with different ideas. Stodolosky, Salk, and Glaessner (1991) found that students characterize experiences in math classes based on their success or ability, whereas students characterize social studies experiences based on interesting or boring topics. Math is one of the most-liked subjects, though as students get older, more students find math difficult and fewer students favor it. Students in NC reported on all subjects.

Reliability and factor structure

The model fit and reliability estimates of our two samples are listed in Table 2, providing evidence of structural validity.

Table 2: Reliability and Model Fit

	Sample 1	Sample 2
Model fit χ^2 (5 *df*)	27.640	59.609
p	<.001	<.001
RMSEA estimate (90% CI)	.05 (.033, .069)	.083 (.065, .103)
CFI	.996	.986
Coefficient α (90% CI)	.80 (.78, .82)	.79 (.77, .82)
Factor loadings		
Item 1	.705	.751
Item 2	.720	.780
Item 3	.624	.497
Item 4	.790	.765
Item 5	.684	.702

Evidence of validity

If our scale is measuring the latent construct of valuing with fidelity, we would expect our measure to correlate more highly with a series of other outcomes *that valuing is theoretically expected to predict*. Using the results from our Sample 2, our valuing scale correlates with Chris Hulleman's scale (which was specifically designed for math) at a high level (.77, .93 corrected for attenuation). Similarly, our scale correlates more strongly with last year's math grades (.23) than last year's non-math grades (.11), and the difference is stat sig (p<.001). Relatedly, valuing correlates more highly with last year's math scores (.24) than last year's science (.18, p=.083), reading (.08, p<.001) or writing (.10, p<.001) scores.

Valuing has a slight, statistically significant negative correlation with age (-0.08), and boys tend to value math more than girls (.26 std. deviations). Consistent with Eccles's (1986) work, which highlights gender differences in boys' and girls' perceptions of math interest and achievement.

Measuring the valuing scale against our other scales has demonstrated a high correlation with the engagement scale, which is class level and related to value (.64), but has weaker correlation with the TSR (.42) and grit (.32) scales, demonstrating that TSR and grit measure distinct things from valuing.

Correlates more strongly with last year's math grades (.23) than Hulleman's scale does (.16) and the difference is statistically significant ($p < .001$).

Correlates more strongly with last year's math scores (.24) than Hulleman's scale does (.18), and the difference is statistically significant ($p = .020$).

Appendix 1: Survey Design Checklist

This set of three checklists is designed for researchers and practitioners with two specific audiences in mind: those who need to choose a pre-existing survey to use and those who need to develop their own instrument. The aim of these checklists is to help survey designers and consumers avoid what are likely to be the largest, and most easily avoidable sources of measurement error. Thus, the lists are not comprehensive, but rather are designed to try to mitigate the largest problems with relatively little effort.

Finally, the checklists are designed as a living document. Your comments about what is helpful/not helpful, clear/opaque, etc. will be invaluable to improving the document. If you think one of the guidelines is off-base, please be sure to include relevant citations in your comments. In addition, information about how you are using the checklists will also be valuable. For instance, are you making decisions about which off-the-shelf survey to use by scoring each one on the checklist and seeing which survey has more "yeses" ticked off? Are you using the checklists to revise previously used surveys? Etc.

Items and response options

Does your survey…	Yes	No
Use scales rather than single items when possible?		
Make sure every item applies to every respondent?		
Avoid item formats consisting of statements and agree/disagree response options…?		
… Instead, use questions and emphasize your focus in your response options?		
Ask one item at a time (thereby avoiding multi-barreled items)?		
Use positive language (because negatives – don't/un/not/etc. – are hard to process)?		
Avoid "reverse-scored" items?		
Choose item formats wisely so that they answer the question you have (e.g., avoid check-all-that-apply)?		
Balance the visual, numeric, and conceptual mid-point of the response options?		

Formatting and ordering surveys

In designing your survey, have you…	Yes	No
Asked the more important items earlier in the survey?		
Labeled each response option?		
Used only verbal labels?		
Visually separated "don't know" and "N/A" response options (e.g., an "I don't know" or "N/A" category)?		
Used only one row or only one column for the response options for each item?		
Ensured that the visual layout of your survey is consistent?		
Placed sensitive questions (e.g., demographics) later in your survey?		

Maximizing responses

In preparing to administer your survey, have you…	Yes	No
Had multiple contacts with your respondents?		
Personalized all correspondences and the survey itself as much as possible?		
Explained how the benefits of taking your survey outweigh the costs?		
Presented the survey as a conversation with your respondents?		
Aligned the stated purpose of your survey with the first item on your survey?		
Strategically and thoughtfully scheduled follow-up contacts with respondents?		

References:

Artino, A. R., Jr., & Gehlbach, H. (2012). AM last page: Avoiding four visual-design pitfalls in survey development. *Academic Medicine: Journal Of The Association Of American Medical Colleges, 87*(10), 1452.

Artino, A. R., Jr., Gehlbach, H., & Durning, S. J. (2011). AM Last Page: Avoiding five common pitfalls of survey design. *Academic Medicine: Journal Of The Association Of American Medical Colleges, 86*(10), 1327-1327.

Artino, A. R., Jr., La Rochelle, J. S., DeZee, K. J., & Gehlbach, H. (2014). AMEE Guide No 87: Developing questionnaires for educational research. *Medical Teacher*. doi: 10.3109/0142159X.2014.889814

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. doi: 10.1037/1040-3590.7.3.309

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*(5), 754-761. doi: 10.1037/0022-006x.56.5.754

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Newbury Park, CA: Sage.

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

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

Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A. (2011). Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. *Social Psychological and Personality Science, 2*(2), 174-181. doi: 10.1177/1948550610385872

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment, 91*(2), 166-174. doi: 10.1080/00223890802634290

Dweck, C. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology, 31*(4), 674-685.

Dweck, C. S. (2006). Test your mindset. from http://mindsetonline.com/testyourmindset/step1.php

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299. doi: 10.1037/1082-989x.4.3.272

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage Publications.

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), 380-387. doi: 10.1037/a0025704

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*, 537-567. doi: 10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed.). Bingley, UK: Emerald Group Publishing.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066X.50.9.741

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. doi: 10.1037/0022-0663.82.1.33

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational & Psychological Measurement, 53*(3), 801-813.

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.

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