PANORAMA EDUCATION

Validity Brief: Panorama Student Survey

This brief outlines the process used to develop the *Panorama Student Survey* and presents research on the reliability and validity of the scales on the survey. We describe evidence from two large-scale administrations of the survey, as well as smaller, targeted studies.

Background

In August 2014, Panorama Education released the *Panorama Student Survey.* From the beginning, Panorama Education has offered the *Panorama Student Survey* as a free and open-source survey instrument through the company website (https://www.panoramaed.com/panorama-student-survey). Panorama Education works with schools, districts, and charter networks to administer the *Panorama Student Survey* to their respective student bodies and provides customized analytics through interactive reports.

This Validity Brief outlines the process used to develop the survey and the results of two large-scale administrations of the survey, as well as specific smaller studies conducted on particular survey scales. These studies indicate that the *Panorama Student Survey* scales have high levels of reliability and demonstrate strong evidence of validity.

The brief is structured to provide guidance about what criteria are important in evaluating survey quality (see the references for more extensive reading on key topics). This report also details key characteristics of the *Panorama Student Survey*, its development process, and results from two administrations so that readers can evaluate these characteristics objectively. As we conduct further studies, Panorama Education will update and add to this Validity Brief. Feel free to contact the Panorama Education Research Team (research@panoramaed.com) to see if there is a more up-to-date version.

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

- Educator-focused design, including survey scales that equip teachers with feedback they can use to improve practice and enable educators to monitor student attitudes, beliefs, and values that are predictive of important outcomes;
- Theoretically-grounded, empirically-based design process that meets or exceeds standards of academic scholarship;
- Adherence to best practices in survey design;
- Allowing schools to customize the survey to their specific needs and teaching frameworks while retaining validity and reliability; and
- Providing the survey instrument to any educator interested in improving pedagogical practice and student outcomes for free.

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 of the Panorama Student Survey

We selected the content for the *Panorama Student Survey* to address the multifaceted needs of teachers, schools, and districts. Specifically, teachers need feedback on their areas of strength, so that they might be further leveraged in the classroom. Equally important, teachers need to identify areas that they can target 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 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. For example, a district that faces challenges with absences and truancy might prioritize the S*chool 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.

The current scales include students' perceptions of:

- **Classroom Climate** – the overall feel of a class including aspects of the physical, social and psychological environment;
- **Engagement** – their 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 Relationship** – the overall social and academic relationship between students and their teachers; and
- **Valuing of the Subject** – how interesting, important, and useful a particular school subject seems.

Survey Development Process: Six Key Steps

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 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 scales were subjected to large-scale pilot tests.

The second important part of the development process 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). Numerous surveys used by educators unfortunately fail to adhere to these well-established survey design practices. For example, designing survey items as statements, particularly ones that require respondents to agree or disagree, are likely to inject additional measurement error into responses. 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, and using too few response options, are 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. The items on the *Panorama Student Survey* adhere to these best practices, which was confirmed during the expert review step.

Statistical Properties and Evidence of Validity

Before describing the psychometric properties, that is, the details of how well these scales measure the psychological attributes they are intended to measure, 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, and may be used with confidence for specific purposes, in specific contexts, and for specific populations.

Pilot Samples

Our main samples are from distinct schools and school districts in 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. 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

Three Main Properties: Reliability, Structural Validity, and Convergent/Discriminant Validity

In the two large-scale pilot administrations, we sought to analyze and measure three main properties of the survey: reliability, structural validity, and convergent/ discriminant validity. Reliability is the property related to whether the item will consistently elicit similar results under similar conditions, so that differences in responses can be attributed to differences in perceptions. Structural validity looks at the extent to which the items of each scale measure one underlying factor or multiple factors. Convergent/ discriminant validity ascertains whether scales designed to measure the same underlying topic correlate highly, while those that measure distinct domains have lower correlations. More technical descriptions of the properties are below.

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 full validity report, the estimates for coefficient alpha for every scale is .70 or greater.

Structural Validity

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.

Convergent/Discriminant validity

In the section on validity, 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 within 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.

Summary:

- Typically, a ratio of .70 or greater is considered adequate reliability for a survey scale (DeVellis, 2003).
- Scales that have undergone confirmatory factor analysis have been subjected to 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 which scales should correlate with a target measure more highly than others.

Validity Evidence for Pedagogical Effectiveness

In this section, we describe the results from the pilot administration for the Pedagogical Effectiveness items of the *Panorama Student Survey*. Validity evidence is available for each of the 10 scales of the *Panorama Student Survey*. To review the others, please contact the Panorama Research Team (research@panoramaed.com). We present the validity evidence for Pedagogical Effectiveness for two reasons. First, we expect that many or most educators who use the *Panorama Student Survey* will elect to use this scale as it directly focuses on students' perceptions of teaching and learning. Second, the validity evidence for the Pedagogical Effectiveness scale attempts to address the same underlying topic that most teacher observation protocols attempt to assess. Thus, correlating student perceptions and administrator observations provides a particularly valuable test of the validity of this measure.

Basic Descriptive Statistics

The table on the next page displays basic descriptive statistics for the Pedagogical Effectiveness items of the *Panorama Student Survey* from the pilot administrations. With both Sample 1 and 2 each item shows substantial variability and moderately strong correlations between each of the items. The correlations are in the expected directions given that, with both samples, items 9-13 were worded negatively. (As an important footnote, these items have since been reworded positively to improve measurement properties.)

Table 2: 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/Discriminant 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).

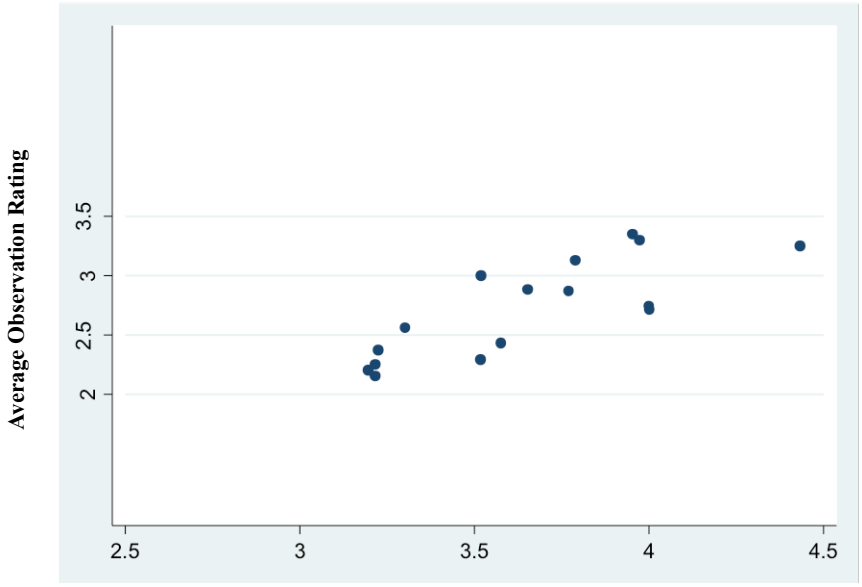
Similarly, the CCSR scale designed to measure distractions in a given classroom was most highly correlated with our item about the teacher's ability to prevent students from getting "out of control" (r = .57) and to prevent time from being wasted (r = .45).

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 Pedagogical Effectiveness survey scale for each of their teachers, usually 5. Those scores were then 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:



In this study, we found students' perceptions and administrators' observations were highly congruent with each other. This suggests that this scale measures pedagogical effectiveness in a way that is similar to administrator observations, which is a particularly strong sign that this scale measures pedagogical effectiveness with a high degree of fidelity.

Conclusion

As with the Pedagogical Effectiveness scale, we have accumulated substantial evidence of validity for the other scales of the *Panorama Student Survey*. Our two large-scale administrations and other studies represent an impressive and growing body of evidence that the *Panorama Student Survey* scales are robust and actionable for districts and schools. We are excited about the promise of these measures to help schools make more informed decisions about their professional development needs and growth.

For more information and to see the full validity report, please contact the Panorama Research Team (research@panoramaed.com).

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