PANORAMA EDUCATION

Reliability and Validity of the Panorama Equity and Inclusion Survey

Students' in-school experiences of diversity, equity, and inclusion matter, and existing measures of their experiences are lacking. This report demonstrates that the Panorama Equity and Inclusion Survey developed by Panorama Education captures substantial and meaningful variability across schools in students' experiences and perceptions, exceeds agreed-upon standards of reliability, and demonstrates strong evidence of multiple types of validity.

Introduction

Districts and schools across the country are increasingly focused on creating school communities that ensure equity and foster inclusion. By asking students to reflect on their experiences of equity and inclusion in school, education leaders can gather actionable data to understand and improve the racial and cultural climate on campus.

The Panorama Equity and Inclusion Survey—developed in partnership with the **RIDES (Reimagining Integration: Diverse & Equitable Schools) Project** at the Harvard Graduate School of Education—provides schools and districts with a clear picture of how students are thinking and feeling about these critical topics. The survey can help schools and districts track the progress of equity initiatives through the lens of student experience, identify areas for celebration and improvement, and signal the importance of equity and inclusion to students, educators, families, and community members.

In this document, we describe how we developed the Panorama Equity and Inclusion Survey and the evidence for its reliability and validity.

Survey Development

To develop this survey, we reviewed existing literature and instrumentation, consulted extensively with scholars and practitioners, followed best practices in the science of item design, conducted a large-scale pilot of the initial instrument, and refined the survey scales based on exploratory factor analyses (see Fowler, 2013; Gehlbach & Artino, 2018; Gehlbach, & Brinkworth, 2011).

Content Creation

After reviewing existing survey instruments and relevant scholarly literature on issues of diversity, equity, and inclusion in education, we—a mix of experts in education, survey methodology, and diversity, equity, and inclusion—spent several months brainstorming questions. Throughout this process, we solicited and received multiple rounds of feedback from practitioners to ensure that the instrument reflected their perspectives. Based on this feedback, we decided to focus primarily on issues of racial diversity, equity, and inclusion because of a) the need to bound the scope of the instrument and b) the primacy of this topic in education and society. This focus is not meant to discount the importance of other forms of diversity, equity, or inclusion, such as gender, sexual orientation, ability, or religion. Instead, we hope our effort demonstrates why measuring students' perspectives on diversity, equity, and inclusion matters, and how to do so.

In developing the survey items, we faced several challenges. First, we needed jargon-free language that students in grades 6-12 could easily understand. We avoided terms such as "equity," "heritage," "oppression," or "cultural competence" in favor of terms that these students would understand. Second, we avoided questions that measured—or even had the appearance of measuring—students' political beliefs. Questions such as, "How important is race in determining who is successful and who is not?" arguably measure political opinions or sociocultural beliefs instead of school experiences, and as such are not appropriate for surveys administered at scale in schools. Third, we needed survey questions that were broadly, if not universally, applicable. Questions such as, "How fairly do students at your school treat people of color?" make little sense for an all-minority or all-white school. Fourth, we needed to minimize potential self-presentation or impression

management biases. Questions such as, "How fairly do you treat students of different races, ethnicity, or cultures?" would likely prompt respondents to give biased answers. Finally, we designed questions for which "higher" responses clearly signaled a healthier racial climate at the school. A "high" response to the question, "How often do adults at your school talk with students about race-related issues?," for example, might reflect a high number of problematic race-related issues that require adult intervention rather than a high degree of proactive conversations about race. In addition, schools with "high" responses to the question, "How often do teachers encourage students to learn about people from your cultural background?" might not identify schools that value diversity, equity, and inclusion—they could just be mostly-white schools.

In creating items, we adhered to the science of survey design best practices (Artino & Gehlbach, 2012; Artino, Gehlbach, & Durning, 2011; Dillman, Smyth, & Christian, 2014; Fowler, 2013; Krosnick, 1999a). For example, we phrase items as questions instead of statements, verbally label all response options, and link response options to each item's underlying concept (Gehlbach & Artino, 2018; Dillman et al., 2014; Krosnick, 1999b; Saris, Revilla, Krosnick, & Shaeffer, 2010; for more, see **Panorama's Survey Design Checklist**). Unfortunately, numerous surveys used by educators fail to adhere to these well-established survey design practices. As a survey scale violates increasing numbers of these best practices, the amount of measurement error and response bias grow.

Through this process, we created an initial set of 28 items that followed best practices in item design and covered a broad range of relevant student experiences and perceptions.

Practitioner Feedback

Next, we received feedback on these new questions from practitioners during the RIDES 2018 "Beyond Desegregation" conference. During a focus group with approximately 20 conference attendees, all of whom were practitioners focused on issues of educational equity, we received specific suggestions on how to improve item phrasing and the instrument. For example, practitioners found the phrase "cultural backgrounds" too vague, and the use of different phrases in different questions (e.g., "cultural background" versus "different race or ethnicity") confusing. This feedback led us to develop a standard phrase across all questions ("different races, ethnicities, or cultures").

In line with Gehlbach and Brinkworth's (2011) recommendation, we also gathered feedback on the survey content from eight education professionals about the relevance of each item. Some of these educators had overseen the pilot data collection in their schools or districts, and all were experts on issues of diversity, equity, and inclusion in education. Their job titles included Equity Program Director, Director of DEI, Assistant Superintendent, Principal, and Director of Community Development. We asked these experts to "tell us how relevant each question is for understanding students' experiences of diversity and equity," and to rate each item as "not at all," "slightly," "quite," "very," or "extremely relevant." In free-response questions, we asked them to indicate "any concerns or ideas about how to improve this item" and "any essential aspects of diversity and equity that we're missing with this set of questions, or any other thoughts you have on how we could improve this survey." Based on this expert feedback, we eliminated two items from the pilot survey instrument.

Pilot Data Collection

We recruited pilot schools and districts through Panorama's and RIDES's educator networks, with the goal of finding educators with a passion for advancing the equity and inclusion climates of their schools. The participating sites were diverse in their geography, size, and student demographics.

In addition to the newly-created survey items, we also administered Panorama's Sense of Belonging scale as part of this pilot. For evidence of its validity, see our validity reports for the **Panorama Student Survey** and the **Panorama Social-Emotional Learning Survey**.

We collected survey responses from 11,679 students enrolled in 22 public middle and high schools from six school districts in the Northeast, Midwest, and Southwest. The majority of students (88%) completed all 33 survey items, and 99% of students skipped four or fewer items. We excluded data from the 42 students (0.3%) who did not complete a majority of survey items.

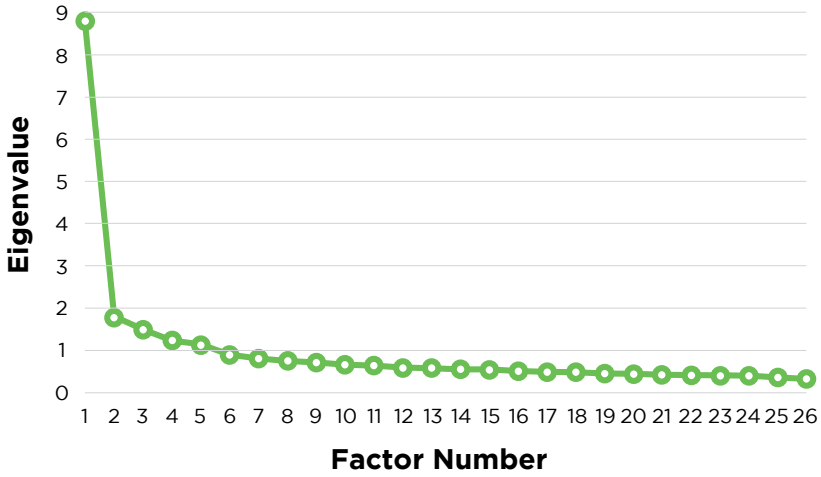
For the vast majority of instances, we used administrative data to determine student demographics; when unavailable, we used available self-report data. Schools ranged in their demographics from 0-92% female (i.e., at least one school was all-male, and one school was 92% female), 0-55% white, 0-82% English-language learners (ELL), and 5-99% eligible for free or reduced-price lunch (FRPL). Overall, the sample was 54% female, 14% ELL, 40% FRPL-eligible, 33% White, 31% Hispanic or Latino, 20% Asian, and 11% Black or African American.

Exploratory Factor Analyses

We conducted exploratory factor analyses on one randomly-selected half of the data (stratified by school), and reserved the other half for confirmatory factor analyses. To verify that these survey data were appropriate for factor analysis, we examined the item intercorrelations, conducted Bartlett's test of sphericity, and calculated the Kaiser-Meyer-Olkin *(KMO)* value. All indicators revealed that the items correlated with each other at sufficiently high and significant levels (mean r = .31), with evidence of underlying latent factors (Bartlett's $p < .0001$; *KMO* = .94).

We conducted exploratory factor analysis with maximum likelihood extraction (see Fabrigar, Wegener, MacCallum & Strahan, 1999) and Promax rotation (see Thompson, 2004) based on best practices and an expectation of correlated factors. Based on a priori considerations, the scree plot (see Figure 1), and a parallel analysis that simulates eigenvalues for a randomly constructed dataset (Horn, 1965; see also Ledesma & Valero-Mora, 2007), we initially decided to extract 2-5 factors.

Figure 1. Scree plot



We removed eight items based on convergent evidence from these initial models. Specifically, we removed items that cross-loaded substantially with other factors (with the difference between the largest and next-largest loading less than .30 in the pattern matrix) for most or all the models.

Applying the same criteria, we iteratively reconducted the exploratory factor analyses with the remaining items, reconsidering the eigenvalues, scree plot, and parallel analysis to determine how many factors to extract, and removing items with substantial cross-loadings. In the second iteration, we extracted two to four factors and removed three items. In the third iteration, we extracted two factors and removed two items.

Before finalizing the scales, we reconsidered excluded items based on their empirical fit with the remaining items and their theoretical importance. First, we added excluded items one at a time to the remaining items, reconducted the factor analysis, and determined which items did not substantially cross-load with the final two factors. Two items, both rated as highly-relevant by experts, passed this test and were added back into the instrument. We removed one item based on its redundancy with a remaining item and relatively low rating of relevance by expert practitioners.

Finally, we faced the choice of what to name the two statistically-derived factors. In other words, what constructs do they capture, and how do we define those constructs? After scrutinizing the items that comprise each topic, reviewing the diversity, equity, and inclusion literature in education, and consulting with expert practitioners, we decided on the below topic names and topic descriptions:

Item A: Cultural Awareness and Action (CAA). *How often students learn about, discuss, and confront issues of race, ethnicity, and culture in school.*

Item B: Diversity and Inclusion (DI). *How diverse, integrated, and fair school is for students from different races, ethnicities, or cultures.*

Table 1 presents the final content and factor loadings.

Table 1. Item Factor Loadings

Item	Item Text	Factor A Loading	Factor B Loading
A1	How often do teachers encourage you to learn about people from different races, ethnicities, or cultures?	0.48	0.11
A2	How often do you think about what someone of a different race, ethnicity, or culture experiences?	0.51	0.06
A3	How confident are you that students at your school can have honest conversations with each other about race?	0.55	0.17
A4	At your school, how often are you encouraged to think more deeply about race-related topics?	0.79	-0.05
A5	How comfortable are you sharing your thoughts about race-related topics with other students at your school?	0.55	0.07
A6	How often do students at your school have important conversations about race, even when they might be uncomfortable?	0.81	-0.14
A7	When there are major news events related to race, how often do adults at your school talk about them with students?	0.71	-0.08
A8	How well does your school help students speak out against racism?	0.58	0.14
B1	How often do you spend time at school with students from different races, ethnicities, or cultures?	0.01	0.64
B2	How often do you have classes with students from different racial, ethnic, or cultural backgrounds?	-0.09	0.71
B3	At your school, how often do students from different races, ethnicities, or cultures hang out with each other?	-0.03	0.81
B4	At your school, how common is it for students to have close friends from different racial, ethnic, or cultural backgrounds?	0.00	0.80
B5	How fairly do students at your school treat people from different races, ethnicities, or cultures?	0.12	0.55
B6	How fairly do adults at your school treat people from different races, ethnicities, or cultures?	0.10	0.47

Note. Factor loadings are from the pattern matrix, with shading indicating their size/direction. Topic A = Cultural Awareness and Action (CAA), Topic B = Diversity and Inclusion (DI).

The extracted factors are psychometrically sufficient in terms of the number of items (min = 6), communalities (min = .29), loadings (min = .47), and cross-loadings (max = .17). They correlate with each other at r = .54. Table 2 presents the item intercorrelations.

Table 2. Item Intercorrelations

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	B4	B5	B6
A1		0.29	0.27	0.4	0.25	0.32	0.37	0.37	0.3	0.19	0.24	0.28	0.25	0.24
A2	0.29		0.32	0.46	0.37	0.37	0.35	0.34	0.3	0.19	0.25	0.29	0.26	0.19
A3	0.27	0.32		0.43	0.53	0.4	0.34	0.4	0.25	0.19	0.32	0.33	0.36	0.29
A4	0.4	0.46	0.43		0.42	0.54	0.48	0.48	0.26	0.16	0.25	0.27	0.29	0.23
A5	0.25	0.37	0.53	0.42		0.39	0.34	0.38	0.25	0.19	0.26	0.28	0.32	0.22
A6	0.32	0.37	0.4	0.54	0.39		0.48	0.47	0.17	0.09	0.19	0.24	0.22	0.14
A7	0.37	0.35	0.34	0.48	0.34	0.48		0.53	0.24	0.16	0.24	0.28	0.26	0.24
A8	0.37	0.34	0.4	0.48	0.38	0.47	0.53		0.26	0.15	0.27	0.32	0.33	0.3
B1	0.3	0.3	0.25	0.26	0.25	0.17	0.24	0.26		0.53	0.46	0.46	0.27	0.24
B2	0.19	0.19	0.19	0.16	0.19	0.09	0.16	0.15	0.53		0.45	0.45	0.26	0.27
B3	0.24	0.25	0.32	0.25	0.26	0.19	0.24	0.27	0.46	0.45		0.64	0.43	0.31
B4	0.28	0.29	0.33	0.27	0.28	0.24	0.28	0.32	0.46	0.45	0.64		0.43	0.33
B5	0.25	0.26	0.36	0.29	0.32	0.22	0.26	0.33	0.27	0.26	0.43	0.43		0.46
B6	0.24	0.19	0.29	0.23	0.22	0.14	0.24	0.3	0.24	0.27	0.31	0.33	0.46	

Note. Cell values are Spearman correlations, with shading indicating their size. Topic A = Cultural Awareness and Action (CAA), Topic B = Diversity and Inclusion (DI).

Reliability

Reliability, as assessed through coefficient alpha, is a measure of signal-to-noise (DeVellis, 2016). In both samples and overall, the two scales demonstrated "good" reliability, and exceeded the typical sufficiency threshold of .70. Table 3 presents the reliability results.

Table 3. Scale Reliability (Cronbach's Alpha)

Sample	CAA	DI
Exploratory	.83	.85
Confirmatory	.83	.84
Overall	.83	.84

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 for specific purposes, in specific contexts, and for specific populations.

Structural Validity

To address structural validity (Messick, 1995), we show evidence of model fit through confirmatory factor analysis results (specifically, comparative fit indices and root mean square error of approximation). Confirmatory factor analysis allows us to determine whether a set of items measures a particular number of constructs. For example, we can test whether items from a single scale measure a single construct (as opposed to multiple constructs). Because confirmatory factor analyses are hypothesis tests, they are more rigorous assessments of structural validity than exploratory factor analyses.

We conducted confirmatory factor analyses on the half of the dataset we did not use for the exploratory factor analyses. We conducted a few different analyses, each getting at a different hypothesis. In one pair of analyses, we segregated items from each topic and separately tested a one-factor model on each scale. In another pair of analyses, we analysed the full confirmatory dataset and tested both a one- and two-factor model. As shown in Table 4, there is sufficient evidence that each scale measures only one construct, and that the items collectively measure two constructs.

Table 4. Results from Confirmatory Factor Analyses

	1-Factor Solution (separate analysis)		1-Factor Solution (combined analysis)	2-Factor Solution (combined analysis)
Statistic	**CAA**	**DI**		
χ^2	805 (*df* = 20)	1063 (*df* = 9)	8305 (*df* =77)	3104 (*df* = 7)
p	<.001	<.001	<.001	<.001
RMSEA (90% *CI*)	0.08 (0.08-0.09)	0.14 (0.14-0.15)	0.14 (0.14-0.14)	0.08 (0.08-0.09)
CFI	0.94	0.91	0.70	0.89

Convergent/Discriminant Validity

To investigate whether the scales showed evidence of convergent and divergent validity, we tested the following hypotheses—all of which are essential for validating scales of this kind:

1. Schools differ substantially on the Equity and Inclusion scales.
2. The Equity and Inclusion scales correlate more with each other than they do with the **Social-Emotional Learning (SEL) scales**.
3. The Equity and Inclusion scales correlate more with more-related SEL constructs and less with less-related constructs.
4. More racially-diverse schools score higher on the Equity and Inclusion scales, particularly Diversity and Inclusion.

Schools differ substantially on the Equity and Inclusion scales. Because these instruments are meant to capture meaningful differences among schools in students' experiences of equity and diversity, we expected schools to differ substantially on each survey topic. Confirming these expectations, one-way ANOVAs demonstrated that both scales capture significant and substantial differences between schools: For the CAA topic, schools accounted for 11.2% of the variance in mean scores, $F(21,11615) = 70.0$, $p < .0001$; for the DI topic, schools accounted for 10.2% of the variance, $F(21,11615) = 62.8$, $p < .0001$. These school effects were larger than the typical effects we see with student reports of social-emotional learning and school climate.

The Equity and Inclusion scales correlate more with each other than they do with the SEL scales. Because we administered Panorama's Sense of Belonging scale along with the piloted survey items, we could examine correlations between CAA, DI, and Sense of Belonging. In addition, the largest participating district administered the following SEL scales: Emotion Regulation, Grit, Growth Mindset, Self-Efficacy, Self-Management, and Social Awareness. Table 5 presents these topic intercorrelations.

Table 5. Topic Intercorrelations

	CAA	DI	SB	ER	Grit	GM	SE	SM	SA
Cultural Awareness and Action (CAA)		0.53	0.51	0.31	0.3	0.22	0.32	0.37	0.47
Diversity and Inclusion (DI)	0.53		0.4	0.23	0.25	0.2	0.23	0.44	0.41
Sense of Belonging (SB)	0.51	0.4		0.42	0.32	0.2	0.39	0.34	0.45
Emotion Regulation (ER)	0.31	0.23	0.42		0.42	0.24	0.46	0.5	0.47
Grit	0.3	0.25	0.32	0.42		0.26	0.53	0.49	0.45
Growth Mindset (GM)	0.22	0.2	0.2	0.24	0.26		0.29	0.28	0.3
Self-Efficacy (SE)	0.32	0.23	0.39	0.46	0.53	0.29		0.49	0.41
Self-Management (SM)	0.37	0.44	0.34	0.5	0.49	0.28	0.49		0.69
Social Awareness (SA)	0.47	0.41	0.45	0.47	0.45	0.3	0.41	0.69	

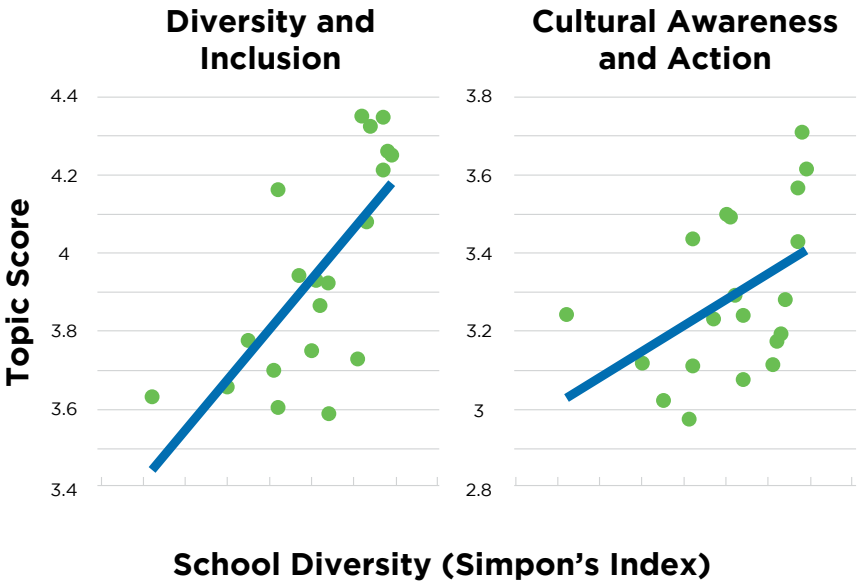
Note. All correlations are significant at $p < .001$. For correlations between CAA, DI, and SB, sample sizes range from 11,603-11,637. For all other correlations, sample sizes range from 3,884-3,890. Shading indicates the size of the correlations.

As expected, the correlation between CAA and DI ($r = .53$) was greater than any correlations between either of these topics and the SEL topics. It was also substantially and significantly greater than the average correlation between the CAA/DI scales and the SEL scales (including SB), $r = .37$.

The Equity and Inclusion scales correlate more with more-related SEL constructs and less with less-related constructs. As shown in Table 5, the CAA and DI topics correlated least with the theoretically-unrelated topics of Growth Mindset, Grit, and Emotion Regulation (mean r = .21, 28, and .27, respectively), and correlated most with the theoretically-related topics of Sense of Belonging and Social Awareness (mean r = .46 and .44, respectively).

More racially-diverse schools score higher on the Equity and Inclusion scales, particularly Diversity and Inclusion. To examine the relationship between school diversity and topic scores, we first calculated a diversity index for each school. We used Simpson's (1948) index, which (in this case) computes the probability that two randomly-selected students are from different race/ethnicity groups; it ranges from 0-1, with higher values indicating greater diversity. In calculating the index, we excluded small-sample (n<50) schools and the low-frequency race/ethnicity groups (American Indian or Alaska Native, Native Hawaiian or other Pacific Islander). Both scales correlated significantly with school diversity as shown in Figure 2, with the correlation substantially higher for DI, r(20) = .71, p = 0.0005, than for CAA, r(20) = .47, p = 0.04.

Figure 2. Correlations between Survey Results and School Diversity



Conclusion

Panorama's new Equity and Inclusion scales align with best practices in survey design and psychometrics. In close collaboration with content experts and educational professionals, we developed two new scales that reliably and validly measure students' experiences of and perspectives on diversity, equity, and inclusion in school.

References

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

Artino Jr, A. R., & Gehlbach, H. (2012). AM last page: Avoiding four visual-design pitfalls in survey development. Academic Medicine, 87(10), 1452.

Artino Jr, A. R., Gehlbach, H., & Durning, S. J. (2011). AM last page: avoiding five common pitfalls of survey design. Academic Medicine, 86(10), 1327.

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

Dillman, D., Smyth, J., & Christian, L. (2014). Chapter 10: Mail questionnaires and implementation. Internet, Phone, Mail, and Mixed-Mode Surveys. The Tailored Design Method. Hoboken, NJ: John Wiley & Sons, Inc, 351-397.

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.

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

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

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.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

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

Krosnick, J. A. (1999b). 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.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. Practical Assessment, Research & Evaluation, 12(2), 1-11.

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.

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.

Simpson, E. H. (1949). Measurement of diversity. Nature, 163, 688.

Thompson, B. (2004). Exploratory and confirmatory factor analysis. Washington, DC: American Psychological Association.

Interested in learning more? **Download the Panorama Equity and Inclusion Survey** and contact us at **info@panoramaed.com** to learn how Panorama can help your school or district measure and take action on this data to advance equity and promote student achievement.

Revision date: 10/19