Leaps Student Voice Survey

Spring 2022 Technical Report

The development and validation of a system for understanding young people's learning experiences

Susan Lyons & Sanford Student

Survey development and validation occurred in close collaboration with the data and learning team at Transcend: David Nitkin, Youssef Shoukry, Ali Picucci, & Adam Haywood

May 16, 2022

LYONS ASSESSMENT CONSULTING

Transcend
BUILDING BEYOND THE LIMITS OF SCHOOL DESIGN

Table of Contents

Executive Summary

The purpose of this report is to provide an overview of the technical documentation associated with the design, development, and on-going validation of the Leaps Student Survey System. The Leaps Student Survey System was developed for use by Transcend in its redesign work with schools related to the 10 Leaps for Equitable, 21st Century Learning. The Leaps Student Survey System contains a set of eleven psychometric scales:

- A Leaps Pulse Check Scale that serves as a quick, diagnostic tool to measure learner experiences related to all ten Leaps, and
- Ten Leaps Deep Dive scales aimed at providing deeper understanding and insight related to learner experiences for each Leap.

The current report contains detailed information related to the intended purpose and uses of the Leaps Student Survey System as articulated in a theory of action (Figure 1), the design and development processes, and the methods and results of analyses aimed at gathering validity evidence to support the technical quality of the instrumentation. In summary, validity evidence related to the content, cognitive processes, internal structure, reliability, and relationships with other variables indicates strong support for interpreting and using the Leaps Pulse Check Scale as intended within the Leaps Student Survey System. The current Spring 2022 pilot is expected to provide the additional evidence needed to support a similarly strong argument for the quality of the 10 Leaps Deep Dive scales.

The validation of an assessment system is an on-going process that requires an accumulation of technical evidence over time. As such, this report concludes with planned next steps in the maintenance and improvement of the Leaps Student Survey System.

Introduction and Theoretical Framework

In early 2021, Transcend partnered with Lyons Assessment Consulting to develop a measurement system that would support deeper understanding and analysis of learner experiences within its partner schools. Transcend is a national nonprofit that supports school communities to create and spread extraordinary, equitable learning environments. In 2020, Transcend released the framework of the 10 Leaps for Equitable, 21st-Century Learning, summarized in Table 1. The Leaps framework was developed after conducting a thorough and systematic synthesis of the latest research on the science of learning and development, and is now used to guide Transcend's transformation work with schools.

Table 1. Transcend's 10 Leaps for Equitable, 21st-Century Learning

Leap	Description
High Expectations with Unlimited Opportunities	All learners experience high expectations and have equitable access to many opportunities, enabling them to progress toward their aspirations for themselves, their families, and the community—regardless of the time and support needed.
Whole-Child Focus	Learners engage in experiences that nurture the totality of cognitive, emotional, social, and physical factors that impact their learning, development, character, and overall health and well-being.
Rigorous Learning	Learners use critical thinking skills to make deep meaning of diverse, complex ideas and are assessed on their ability to apply, analyze, and use their knowledge in creative ways across contexts.
Relevance	Learning explores young peoples' interests and goals, is connected to their communities, and enables them to understand and tackle real problems in authentic contexts.
Affirmation of Self & Others	Each learner develops a unique, positive sense of self and purpose as well as a deep respect for the identities of others; these diverse identities are celebrated, nurtured, and leveraged in meaningful and anti-oppressive ways to support everyone's learning.

Social Consciousness & Action	Learners critically examine social problems and work toward a more just world; they develop the knowledge, skills, and mindsets needed to continue taking anti-oppressive actions that disrupt and dismantle racism and other inequities.
Connection & Community	The environment is relationship-rich: learners are deeply known and respected by a variety of adults and peers; collaborate closely; and form meaningful relationships across lines of difference that nurture empathy, foster belonging, support well-being, and build social capital.
Customization	The focus, pace, and sequence of learning, as well as the resources and supports provided, are tailored to each learner's identity, prior knowledge, development, way of learning, and life experiences, ensuring that all learners have what they need to be successful and those who need more receive more.
Active Self-Direction	Young people are active drivers of their learning; they grapple directly with concepts while receiving adult and peer guidance and support; they have a voice in decisions about how and what they learn, so that the process grows agency and meaningfully builds on their interests and prior knowledge.
Anytime, Anywhere Learning	Learning can happen anywhere and at any time for all learners with teachers, families, community members, and other important figures in a young person's life all playing important educational roles.

Lyons Assessment Consulting is a leader in designing innovative assessment systems intended to disrupt systems of oppression and promote social justice. This report details the process and outcomes associated with developing a measurement system that would support Transcend's work with its partner schools using the 10 Leaps framework.

Leaps Student Survey System Theory of Action

High-quality systems of assessment are designed with a well-articulated theory of action that details how the assessments are intended to work together to serve their intended purposes and uses in service to larger programmatic goals. Figure 1 on the next page outlines a high-level theory of action for the Leaps Student Survey System. The goal that the measurement system is aiming to support is articulated At the bottom of the figure in green, the components of the assessment system are indicated in dark blue, and the actions and assumptions associated with system use are indicated in light blue.

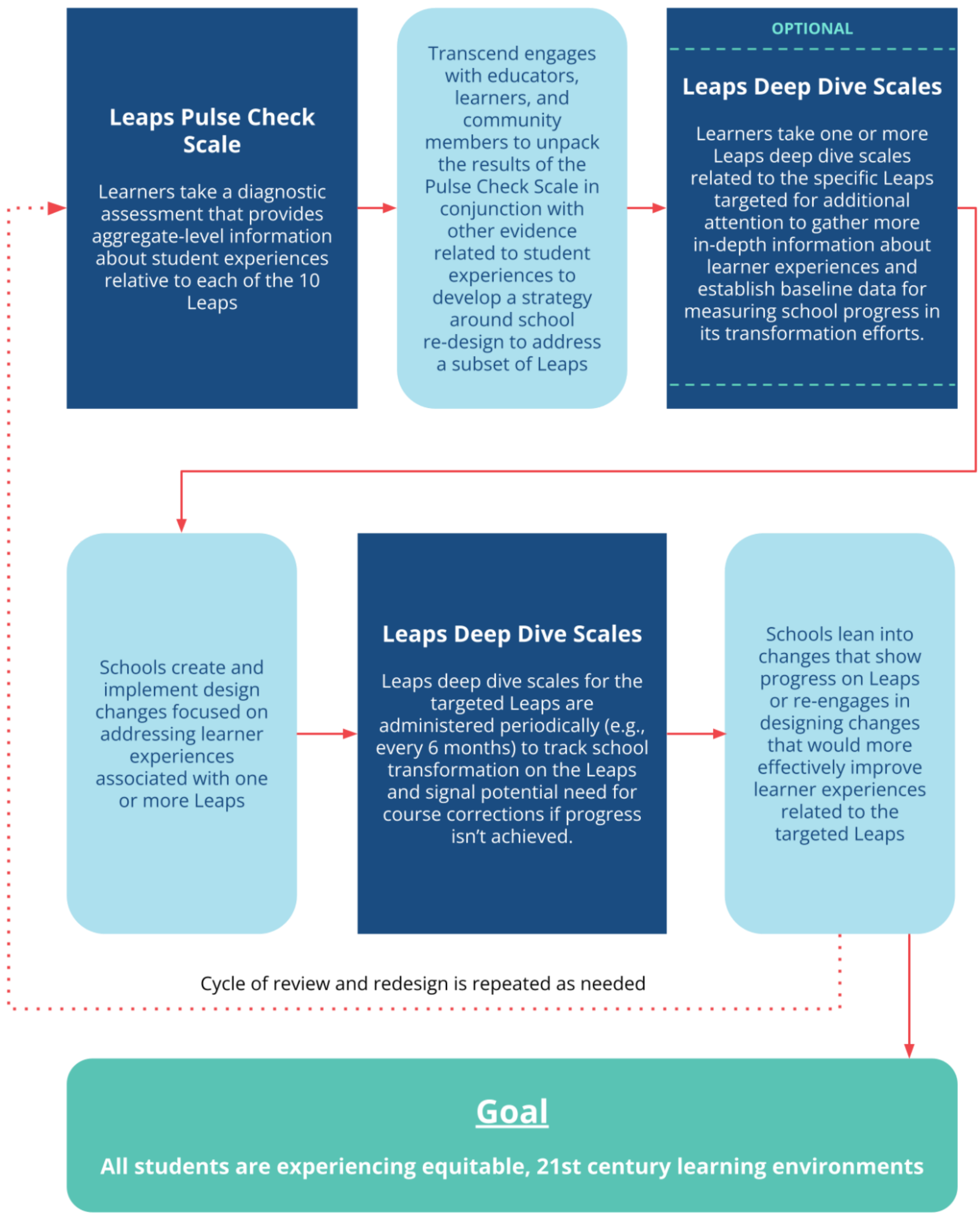


Figure 1. Leaps Student Survey System Theory of Action

Overview of Scale development and Validation

As detailed in Figure 1, the Leaps Student Survey System comprises two primary assessment components:

1. **A Leaps Pulse Check Scale** that captures learner experiences relative to all ten Leaps to provide a quick, diagnostic view of student perceptions and experiences related to all Leaps.
2. **Deep Dive Leaps Scales (10 assessment tools),** each designed to provide a more in-depth measurement of student experiences relative to the Leaps.

In order for the system to operate optimally, the 10 Deep Dive Leaps Scales are intended to be on the same measurement scales as the individual Leaps information derived from the Leaps Pulse Check Scale. This will allow scores to be meaningfully compared across tools and measurement occasions.

The scale development and validation processes completed thus far are summarized in Table 2.

Table 2. Summary of Scale Development and Validation

Timeframe	Activity
Winter 2021	Assessment system conceptualization
	Content expert analysis of each Leap to identify central sub-constructs that comprise the definition of each Leap
Spring 2021	Literature reviews to identify existing scales related to the sub-constructs represented within each Leap
	Identification of items within existing scales that best represent the intended Leap
	Seeking permission to use and/or modify existing items
	Content expert item development and review to modify and write new items where needed
	Cognitive labs with students to review and revise newly-developed items
	Piloting the 10 Deep Dive Leaps Scales with sample of Transcend's partner schools

Summer 2021	Analyzing pilot data to create final Deep Dive Leaps Scales
	Conducting another round of cognitive labs with a subset of scales
	Constructing the Leaps Pulse Check Scale using item statistics and content expert judgment from the Deep Dive Leaps Scales
Fall & Winter 2021-2022	Administering the Leaps Pulse Check Scale with large sample of Transcend's partner schools
	Analyzing administration data and finalizing Leaps Pulse Check Scale
Spring 2022	Writing technical documentation and recommending next steps for on-going improvement and technical maintenance of the Leaps Student Survey System.
Summer 2022	Additional validation and scaling work to be determined based on recommendations provided in this report.

The 10 Deep Dive Leaps Scales

Scale Development

Our scale development approach began with a desire to leverage existing validated scales that target similar constructs to those of the Leaps. The intention was to adopt or adapt as many existing items as possible in order to benefit from the existing validation work that went into the development of other scales. Scales were primarily drawn from the academic literature using a library database search, but also were found through other partner non-profits engaged in similar work with schools. All items used or adapted for our scales are either explicitly open-source or written permission for use was sought from the scale copyright holders.

In order to find scales that best aligned with the intended measurement targets of our Leaps, Lyons Assessment Consulting worked closely with content experts at Transcend to identify a set of sub-constructs that comprise the definitions of each of the Leaps. Not only was this helpful for our literature search, it also served as an important step in more fully operationalizing the definitions of the Leaps for the purpose of measurement. Table 3 shows the sub-constructs identified for each of the 10 Leaps and the source scales from which we were able to adopt or adapt items.

Table 3. Targeted Sub-Constructs that Comprise each Leap

Leap	Sub-Constructs	Source Scales
High Expectations with Unlimited Opportunities	High expectations	– Aldridge & Fraser, 2008 – University of Chicago, 2021 – MCIEA, 2021
	Equitable access to opportunities	
Whole-Child Focus	Support for social-emotional learning	N/A all new items
	Support for physical health	
	Support for mental health	
	Support for identity development	

Rigorous Learning	Critical thinking	– Ferguson, 2010 – Chai et al., 2015
	Task rigor	
Relevance	Attention to learner goals/interests	– Assor et al., 2002 – Burns, 2006 – Chai et al., 2015 – Frymier & Shulman, 1995 – Young et al., 2008
	Connection to the real world	
	Connection to prior learning	
Affirmation of Self & Others	Affirmation in school	– MCIEA, 2021
	Diversity and inclusion	
Social Consciousness & Action	Support for critical consciousness	– Panorama, 2021
Connection & Community	Connections with adults	– Appleton et al., 2006 – Chai et al., 2015 – MCIEA, 2021
	Connections to peers	
	Connection to community	
	Collaboration	
Customization	Personalized support	– Aldridge & Fraser, 2008 – Assor et al., 2002 – Ferguson, 2010 – University of Chicago, 2021 – US Department of Education, 2017
	Personalized speed	
	Personalized materials and resources	
Active Self-Direction	Student choice	– Appleton et al., 2006 – Assor et al., 2002 – Burns, 2006 – Ferguson, 2010
	Empowerment at school	
	Self-directed learning	
Anytime, Anywhere Learning	Learning outside of school building	N/A all new items
	Learning outside of school hours	
	Valuing other sources of learning	

Even with the excellent survey sources from which we drew our initial set of items, for all scales, we nonetheless needed to engage in some new item development. This was exclusively true for all items comprising the Whole-Child Focus and Anytime, Anywhere Learning Leaps. Item development occurred iteratively with rounds of input, review, and revision from the Lyons Assessment Consulting and Transcend teams. We relied heavily on the content expertise of the Transcend learning engineers who were closely involved in the Leaps framework development. Additionally, we drew from existing resources in the field to help shape our understanding about the most central ideas associated with each Leap. For example, to help us define the sub-constructs and ultimately write items associated with the Whole Child Focus, we drew on the Whole Child Framework developed by the Chan-Zuckerberg Initiative (CZI); see Figure 2.

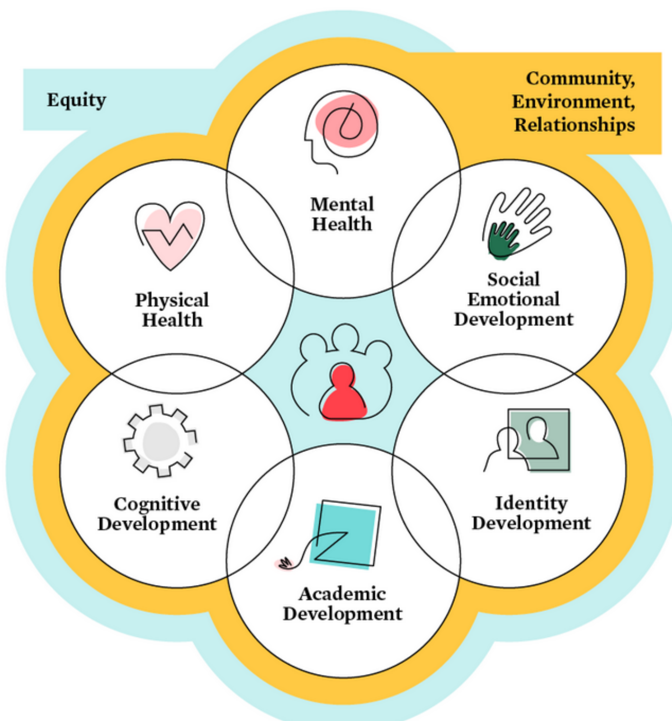


Figure 2. Chan-Zuckerberg Initiative Whole Child Framework

Because multiple Leaps attend to cognitive and academic development, item development for the Whole Child Focus Leap focused on the physical health, mental health, social-emotional development, and identity development aspects of the CZI framework.

All items that were newly developed for our Spring 2021 piloting of the Leaps Deep Dive Scales were tested with students using a Cognitive Laboratory protocol. The purpose of the cognitive laboratory is to gauge the degree to which the items, as written, are eliciting the intended cognitive processes as students read and respond to the scales. Seven students ranging from grades 3-9 participated in these Cognitive Labs in which they read the item aloud and then shared their thinking as they selected their response to the item on the provided Likert scale. In some cases, item wording was difficult to understand or elicited thinking that was not in line with the intended meaning of the item. These cognitive labs were thus highly informative for making item revisions and refinements before we conducted formal pilot testing with a large sample of students.

Pilot Analyses

Administration

Following initial development, each scale was piloted during the spring of 2021. Students across six schools participated in the pilot, with a total of 1670 students responding to at least one Leap scale. Of these students, the large majority were in grades 3-8, although the 42 students from School 3 included in the pilot were Grade 1 students. The results and recommendations associated with administering the Leaps Student Survey System in Grade 1 are discussed further in the sub-section titled "Piloting with younger students." As shown in Figure 3, of the remaining students, grades 6-8 were most heavily represented.

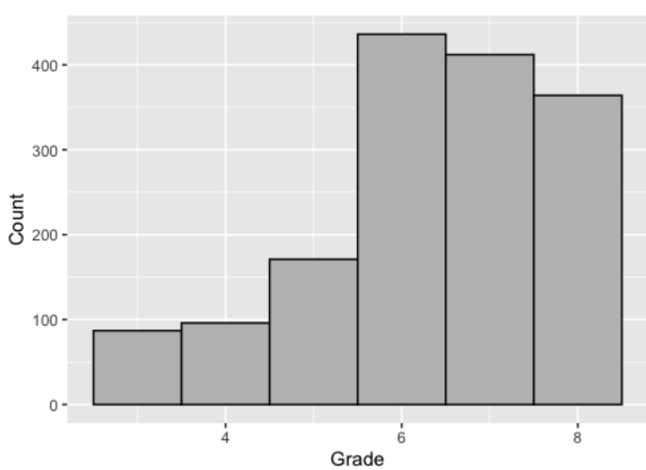


Figure 3. Distribution of students' grades in Leaps pilot sample

Table 4 includes count and demographic summaries for each school including free and reduced priced lunch eligibility (FRPL), special education or individual education program designation (SPED/IEP), gender, English language learner (ELL) status, and the most commonly represented racial/ethnic subgroups. Note that School 1 designated Hispanic/Latino students as both Hispanic/Latino and White; other schools did not employ this practice.

The majority of pilot respondents were students at School 1. Although not all schools provided all (or, in some cases, any) relevant demographic information, most responses came from schools where the majority of students are non-White. Where available, FRLP eligibility is moderate to high, indicating economic as well as racial/ethnic diversity.

Table 4. Demographic summary of pilot schools

School	n	FRLP	SPED/IEP	Female	ELL	Black	Hispanic/ Latino	White
1	1025	0.50	0.14	0.49	0.12	0.09	0.70	0.92
2	62	–	–	–	–	–	–	–
3	42	–	–	–	–	–	–	–
4	324	0.51	0.12*	0.60	–	0.12	0.71	0.16
5	88	0.94	0.18	0.41	0.19	0.45	0.44	0.08
6	129	–	0.06*	0.56	0.03	0.67	0.28	0.02

Note: Schools used either SPED or IEP designations, but not both. Asterisks indicate the use of IEP.

School leaders chose to administer one or more of the Leap Deep Dive scales to their students. Table 5 shows which Leaps scales were administered at each school. Here we can see that although most respondents came from School 1, School 1 only administered the RL and WCF scales. This is why, as shown below in the “N” column of Table 5, the sample sizes vary widely by scale for the pilot analyses. It is also important to note that the CUS scale was only administered at School 3, where respondents were exclusively in Grade 1 and responded to items orally. These

School 3 responses were excluded from initial analyses of the RL scale, and analyses of the CUS scale must be interpreted with caution. Similarly, administration at School 2 did not follow the intended procedure, so responses from School 2 were excluded when responses from another school were available. For Leaps administered only at School 2, analysis results should again be interpreted tentatively. No schools administered the scale associated with the Anytime, Anywhere Learning Leap and therefore no analyses or results are reported. Piloting the scale for this Leap is a priority for Spring 2022.

Table 5. Leap administration by school with sample size

Leap	Scale	School 1	School 2	School 3	School 4	School 5	School 6	n
Active Self-Direction	ASD	–	Yes	–	–	–	–	62
Affirmation of Self & Others	ASO	–	Yes	–	Yes	–	–	385
Connection & Community	CC	–	Yes	–	Yes	Yes	Yes	576
Customization	CUS	–	–	Yes	–	–	–	39
High Expectations & Unlimited Opportunities	HE	–	Yes	–	Yes	Yes	–	403
Relevance	REL	–	–	–	–	–	Yes	129
Rigorous Learning	RL	Yes	Yes	Yes	–	–	–	1059
Social Consciousness & Action	SCA	–	Yes	–	–	–	–	62
Whole-Child Focus	WCF	Yes	–	–	Yes	Yes	–	1361
Anytime, Anywhere Learning	AAL	–	–	–	–	–	–	–

Initial descriptive analyses

Mean responses and standard deviations (SDs) were calculated for all items individually, each scale overall, and disaggregated by student group. Additionally, item-total correlations were calculated for the items on each scale. Item-total correlations were calculated excluding the item being analyzed to avoid inflation of the correlation; item responses are scored on a 1-to-5 scale. All statistics were calculated from complete response strings; respondents who skipped at least one item were excluded. These initial analyses were intended to flag any items unlikely to be useful in operation, due either to extreme response patterns (e.g., most students selecting the highest or lowest level of agreement) or to potentially targeting a different construct (as indicated by a low item-total correlation). The main goal at this point was to identify potential items for removal from each scale, as the original scales were deemed too long for regular operational use and the target final length is between 8-10 items each.

The descriptive analysis results for each scale are detailed in Appendix A. A summary for each scale is offered here.

For the ASD scale, initial item-total correlations were generally strong, and mean item responses did not fall at either extreme of the scale. Of the 15 pilot items, the seven items with the lowest item-total correlations (ASD_Q_1_1, ASD_Q_1_2, ASD_Q_1_3, ASD_Q_1_6, ASD_Q_1_7, ASD_Q_1_8, and ASD_Q_2_2) were initially flagged as candidates for scale reduction to meet our target scale length of 8-10 items.

For the ASO scale, the item statistics presented some evidence of extreme item response patterns for several items where nearly all students responded with a 4 or 5 (the two highest levels of agreement). 15 total items were piloted, and 7 of the items exhibited a mean response above 4 (ASO_Q_1_3, ASO_Q_1_4, ASO_Q_1_5, ASO_Q_2_2, ASO_Q_2_3, ASO_Q_3_1, and ASO_Q_4_2). These items were initially flagged for removal. Item-total correlations were also generally lower for this scale than for other scales, such as ASD. One possible source of low item-total correlations is multidimensionality, and as described in the next section, our analysis of the dimensionality of this scale did lead to further suggested reductions.

Item responses for the CC scale were broadly acceptable, with no indication of extreme responses and generally strong item-total correlations. Three items (CC_Q_1_1, CC_Q_1_3, and CC_Q_1_6) were considered as candidates for removal based on showing the lowest item-total correlations. Items CC_Q_1_11 and CC_Q_1_9 were also flagged for removal due to content-based decisions, producing an 8-item scale at this stage.

Item responses for the CUS scale were not extreme, but item-total correlations were far lower for this scale than for others. Recall that this scale was administered only to students below the intended grade level for the Leaps. Because of this, rather than recommend removal of items based upon low item-total correlations, a second pilot with older students is recommended and planned for April 2022.

Of the 14 items piloted for the HE scale, four items (HE_Q_1_1, HE_Q_4_1, HE_1_2, HE_1_4) were initially flagged due to low item-total correlations relative to the rest of the HE items. This did leave one item with a high mean response of 4.08, HE_Q_1_5, in the suggested revised scale.

For the REL scale, overall item-total correlations were strong and mean responses were not extreme. Of the 13 items piloted, 3 items (REL_Q_2_6, REL_Q_2_2, REL_Q_2_5) were considered for removal to reduce the length of the scale to ten.

Responses to the RL scale items were not extreme and item-total correlations were generally strong and consistent. Item RL_Q_1_8 had the lowest item-total correlation, so it was initially flagged for potential removal to reduce scale length to ten.

Because the SCA scale was already shorter than ten items and no items displayed extreme response patterns, no items were flagged for removal at this stage. SCA_Q_1_2's item-total correlation was lower than others', but not to the point that it was deemed to be problematic.

Responses to WCF scale items were not extreme, and item-total correlations were consistently strong. With the scale already shorter than ten items, no items were flagged for removal at this stage.

Analysis of internal structure: Exploratory factor analyses

For each scale, item responses were analyzed using exploratory factor analysis (EFA). EFA is a method of analyzing the covariance of items on a given scale to understand the extent to which they appear to reflect the same underlying construct(s). For Leaps scales, a unidimensional solution--that is, a solution in which all items load on a single construct--is desirable because it supports the choice to use a single scale score, sum or average score to summarize a given student's responses to the scale.

In general, visual scree tests (Cattell, 1966) indicated the presence of a strong first factor for all scales, corroborated in most cases by the Kaiser-Guttman test (Kaiser, 1960). This justifies the choice to treat most scales as unidimensional for the purposes of summarizing across item responses. However, the EFA of the ASO scale indicated two coherent factors. This was incorporated into the scale-shortening recommendations described in the next section. Appendix B includes scree plots for each scale, as well as tables including variance explained by up to four factors. Note that the sample sizes for the ASD, SCA and CUS scales were not sufficient to perform an EFA; instead, we relied on the descriptive analyses and initial reliability analysis to make final scale recommendations for these Leaps.

Reliability analyses

As noted above, one goal of pilot testing was to inform the reduction of each scale to fewer items. We began by calculating Cronbach's α for each complete scale. The next step was to recommend a strategy for shortening each scale. For scales with more than ten original items that appeared unidimensional, recommendations were made to reduce the scale to at most ten items by removing the items with the lowest item-total correlation and/or factor loading.

Initial reliability coefficients are found in Table 6. These indicate that most full scales' reliability at the individual level is quite high; higher, in fact, than required for a scale that will used to primarily support aggregate (e.g., class-level) inferences.

Table 6. Cronbach's α for each pilot scale

Leap	Cronbach's α
ASD	0.95
ASO	0.85
CUS	0.73
HE	0.86
RL	0.88
WCF	0.88
CC	0.93
REL	0.91
SCA	0.89

For most scales, the recommended final scale length was 8-10 items long. For the ASO scale, the emergence of a clear second factor led to the recommendation to shorten the scale to just the items that loaded most strongly on this second factor. The two primary factors could be interpreted as:

1. Racial/ethnic diversity and inclusion of the student body
2. Cultural representation and relevance within the instruction and curriculum

The second factor was deemed to be more substantively related to the purpose of the Leap than the first factor, and was therefore retained at the exclusion of the items loading on the first factor. The five remaining items, loading on the second factor, are provided in Table 7.

Table 7. Five items loading on ASO Factor 2

Item ID	Item Stem
ASO_Q_1_1	In your school, how often do you see people like you represented in what you study?
ASO_Q_1_2	In your school, how often do you see many different kinds of people represented in what you study?

ASO_Q_2_1	My school provides instructional materials (e.g., textbooks, handouts) that reflect my cultural background, ethnicity, and identity.
ASO_Q_5_1	How valued do you think all students' home cultures and languages are in your school's curriculum?
ASO_Q_5_2	How valued do you think your home culture and language are in your school's curriculum?

Differential item functioning analyses

As noted above, item-level means and SDs were disaggregated by student demographic variables during the initial descriptive analyses. However, these analyses alone cannot detect item-level bias; instead, they can reflect bias, real group-level differences in the underlying construct, or a combination of the two. Measurement invariance analyses are intended to disentangle these two potential sources of group-level differences in item responses to produce evidence that any variability in a given item's ability to elicit the intended construct is random and unassociated with students' backgrounds. For example, non-invariance across two racial groups–for example, Black and White identifying students–for an item would indicate that the item cannot be interpreted as comparable across these student groups due to meaningful differences in how the students are interacting with this item. This would mean that responses on the scale cannot be interpreted and used as intended.

Our analysis of measurement invariance for each scale focused primarily on differential item functioning (DIF). For each item, DIF analysis looks for evidence that respondents of similar overall levels on the underlying construct differ systematically in their responses to the given item according to known demographic variables. An example of DIF would be a situation in which Black students with similar overall scores to White students on the WCF scale tended to respond substantially lower to a specific item on the scale. If the difference is large enough, this situation would have the potential to attenuate group-level summaries of responses not only to the item itself, but to the scale as a whole, potentially misleading stakeholders attempting to improve whole-child strategies in their classroom or school.

There are many methods for detecting DIF; we selected logistic ordinal regression DIF analysis. Logistic regression DIF analysis was selected as the DIF detection method for this analysis because it is easily implemented for polytomous item responses such as ours, as outlined in Zumbo (1999). Logistic ordinal regression DIF analysis looks for DIF by fitting a logistic regression model where the outcome variable is correctly responding to an item and where the predictor variables are group membership, sum score across all items of the scale, and an interaction of the two.

Logistic regression DIF analysis consists of up to three regression models per item. The first regresses responses to a given item on examinees' total score. The second adds the grouping variable, for example gender, to the analysis. The third adds an interaction term to the model. A chi-square test of the third model against the first with two degrees of freedom then serves as a simultaneous test of uniform and non-uniform DIF. One can also test the second model against the first as a test of just uniform DIF. Zumbo (1999) contains a great deal of information on this method. For this analysis, our focus was on the simultaneous test of uniform and non-uniform DIF, which provides the most information on the presence of any DIF for a given item. An item was flagged for statistically significant DIF if the chi-squared test of differences produced a p-value less than 0.05. This is a very generous p-value considering the number of hypothesis tests conducted; it was selected out of an abundance of caution to ensure that even borderline DIF could be detected.

Statistical significance is only part of the story in logistic regression DIF analysis; one must then calculate DIF "effect sizes" to help decide if the DIF is meaningful or not. To do this, one must look at the magnitude of DIF for all items (per Zumbo's advice to calculate all effect sizes, then make sense of them in tandem with statistical significance). In logistic regression DIF analysis, this is represented by the difference in R-squared between models 1 (item response on total score) and 3 (item response on total score, gender and interaction). Again, in keeping with Zumbo (1999), we calculated the McKelvey-Zavoina pseudo-R-squared measure for each model and differentiate between small and large amounts of DIF with a cutoff of 0.13. This cutoff is somewhat arbitrary and could certainly be stratified further; we treated a difference of less than 0.05 as trivial. Appendix C contains detailed information on statistical significance and effect sizes for all items on all scales.

DIF analysis is only possible when student demographic data are available. Because Schools 2 and 3 did not provide demographic data, we could not conduct DIF analyses for the CUS, ASD or SCA scales. DIF analyses for these three scales are being pursued in the April 2022 pilot. The scales analyzed were REL, CC, WCF, RL, HE and ASO.

For most scales, the group with the largest sample size in the item response dataset was treated as the “default” group- so, for example, if Hispanic/Latino students were most represented in a dataset, then that group was treated as the default against which to compare the responses of students from other groups for DIF. Analyses were conducted for both gender and race/ethnicity. All analyzed scales had sufficient sample size for analysis of potential DIF by gender. For race/ethnicity, analyses were generally only conducted for pairs of race/ethnicity variables with more than 60 students. Details on which analyses were conducted for each scale are available in Appendix C. DIF analysis was conducted using the shortened, 8-to-10-item version of each scale.

Respondent demographics are reported in Table 8.

Table 8. Respondent demographics for each scale

Scale	N	Female	Male	Black	Hisp.	White	Asian	Am. Ind./ Al. Nat.	Nat. Ha./ Pac. Is.
ASO	359	179	118	35	212	46	2	1	0
CC	478	283	195	131	279	60	2	1	1
HE	376	188	126	44	216	49	2	1	0
REL	112	64	48	75	31	3	0	0	0
RL	994	478	487	85	673	889	13	29	3
WCF	1268	662	606	124	875	921	16	29	2

When considering both statistical significance and effect size, there was no evidence of DIF for any item across the groups we were able to analyze. On most scales, there was at least one item displaying statistically significant DIF, but for every single item except one on the REL scale, the effect size of DIF was trivial. These findings and implications for this one item are discussed in detail below.

REL scale

Given limitations related to sample size for this scale, we only analyzed DIF by gender. The items REL_Q_1_1, REL_Q_1_4 and REL_Q_2_1 were flagged for statistically significant DIF. The magnitude of DIF for REL_Q_1_1 was small (0.09). All other DIF effect sizes for the items on this scale were trivially small.

To investigate REL_Q_1_1, we produced empirical item characteristic curves (ICCs) for male and female respondents. Figure 4 shows evidence of non-uniform DIF where female students with lower total REL sum scores were likely to respond less negatively to REL_Q_1_1 than male students with similar overall scores. The phrasing of this item is, "adults at my school help me see how what I am learning relates to my life." In terms of the item's content, there was no clear reason why this item would differ in its meaning for female vs. male students. Given the very large number of comparisons done in the course of our DIF analyses, an occasional spurious finding is expected, and we treated this single instance of small DIF as such.

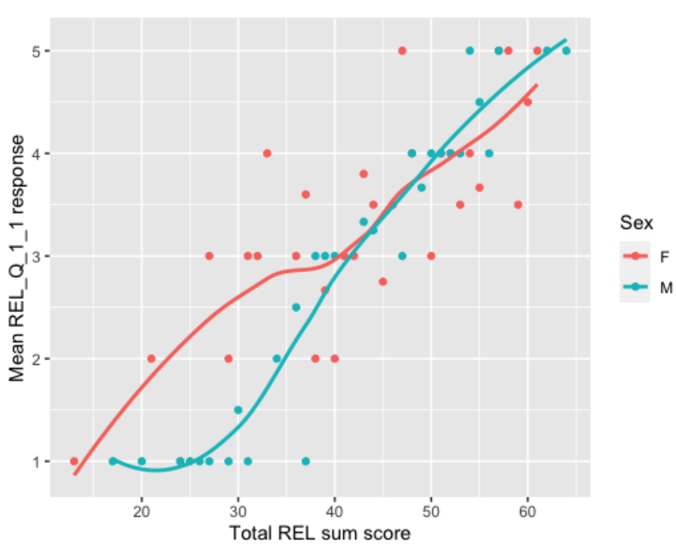


Figure 4. Empirical ICC for REL_Q_1_1 for male and female students.

Piloting with younger students

The Leaps were originally designed for at least a third-grade reading level. However, during piloting, the opportunity arose to pilot the RL and CUS Leaps Deep Dives with students in first grade. The intent here was to explore the potential appropriateness for using these scales with younger students below grade 3.

All analyses were conducted at a single school, School 3. In total, 39 students responded to the pilot versions of two Leaps scales, CUS and RL. Because of the reading requirements of the Leaps, for this Grade 1 pilot, the teachers read each item aloud and recorded each student's response. To analyze the quality of results we compared the younger students' responses on the RL scale to our larger sample of responses on the RL scale. For the CUS scale, administration complications resulted in only students at School 3 responding to this scale. Although it would have been ideal to have a comparison group of older students, we still analyzed the responses of younger students to the CUS scale on their own.

Reliability

Coefficient α for the RL scale administered to School 3 students was 0.57. In contrast, α for the RL scale administered to students in the intended grade range was 0.88. This indicated that the reliability of responses from younger students recorded orally was quite degraded, which is consistent with prior literature.

Although there was no comparison group for the CUS scale, we found that coefficient α was equal to 0.71. This was just high enough to be considered acceptable for the inferences that the Leaps were intended to support, but recall that the other scales exhibited much higher α and could be reduced significantly while maintaining high reliability. For our Grade 1 pilot, α was barely above 0.70 with 14 items.

Item statistics and response distributions

Next, we turned to item-level analyses. First, we looked at the items on the RL scale. Table 9 shows a comparison of the descriptive statistics for our full sample compared to the Grade 1 sample.

Table 9. Descriptive statistics compared across age cohorts

	3-8 Pilot Results				Grade 1 Pilot Results			
Item	N	Mean	SD	Item-total correlation	N	Mean	SD	Item-total correlation
RL_Q_1_1	965	3.53	1.09	0.61	37	3.86	0.48	-0.13
RL_Q_1_2	965	3.39	1.09	0.60	37	3.30	0.94	0.62
RL_Q_1_3	965	3.72	1.09	0.60	37	3.86	0.48	-0.06
RL_Q_1_4	965	3.79	1.02	0.63	37	3.76	0.60	0.51
RL_Q_1_5	965	3.37	1.05	0.66	36	3.67	0.72	0.34
RL_Q_1_6	965	3.51	1.01	0.60	36	3.89	0.40	0.30
RL_Q_1_7	965	3.66	0.97	0.55	36	3.58	0.77	0.43
RL_Q_1_8	965	3.5	1.01	0.54	36	3.28	0.94	0.18
RL_Q_2_1	965	3.04	1.18	0.60	34	3.24	1.05	0.19
RL_Q_2_2	965	3.32	1.19	0.61	30	3.40	0.97	0.23
RL_Q_2_3	965	3.07	1.18	0.59	30	3.33	0.92	0.12

There are two main issues that are immediately noticeable. First, the standard deviations are much smaller for the Grade 1 students than they are for older students. This indicates the possibility that students responding aloud tend to agree more than students who are able to respond individually and anonymously (i.e., without the presence of a teacher). This is also shown in Figure 5, which compares the distributions of mean item responses to the RL scale. We can see that the distribution for K-2 students is noticeably less variable and more positive.

Second, the item-total correlations are much lower for the Grade 1 students, including two negative item-total correlations. This indicates that the RL scale did not function as intended for younger students.

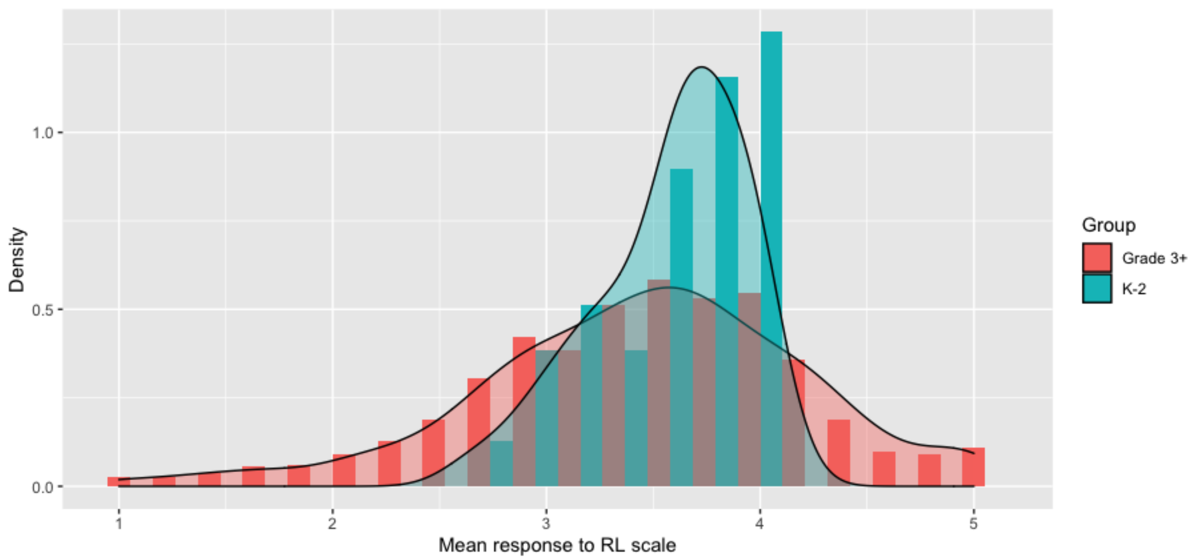


Figure 5. Comparison of response distribution by grade group to RL scale

For the CUS scale, the lack of a comparison group limited our analyses. Shown in Table 10, we were able to calculate item statistics, albeit just for the Grade1 students. The distribution of mean scale response was a bit wider compared to the RL scale for Grade 1 students, however, the item-total correlations were still inconsistent, including a negative item-total correlation.

Table 10. Item Descriptives for Grade 1 Sample

Item	N	Mean	SD	Item-total correlation
CUS_Q_1_1	39	3.21	0.98	0.59
CUS_Q_1_2	39	3.13	1.13	0.34
CUS_Q_1_3	39	3.05	1.07	0.34
CUS_Q_1_4	39	2.87	1.20	0.38
CUS_Q_1_5	39	2.95	1.02	0.31
CUS_Q_1_6	39	3.15	0.96	0.52
CUS_Q_1_7	39	3.26	1.16	0.51
CUS_Q_2_1	39	3.46	0.91	0.04
CUS_Q_2_2	39	3.31	0.89	0.14
CUS_Q_2_3	38	3.66	0.67	0.27
CUS_Q_2_4	36	3.69	0.71	-0.04

CUS_Q_2_5	35	3.54	0.82	0.53
CUS_Q_3_1	36	3.44	0.91	0.16
CUS_Q_3_1	35	3.71	0.71	0.39

These analyses indicated that the responses gathered from Grade 1 students appeared to be less trustworthy and reliable compared to those gathered from older students. As such, we cannot recommend the Leaps Student Survey System for use with students below grade 3.

Revised Leaps Deep Dive Scales

Based on all of the analyses conducted in our initial Spring 2021 pilot, we produced a set of revised 10 Leaps Deep Dive Scales for use with Transcend's partner schools. While the ultimate goal remains to reach a final set of 8-10 items for each scale, the team determined the best course of action was to undergo another round of item generation and testing that retains many of the items from the Spring 2021 pilot and adds additional test items that reflect our new, deepened understanding of the Leaps constructs and measurement approach. Table 11 below provides a summary of the revised scales that are being actively pilot tested in Spring 2021. Final scales will be published and shared in full at the conclusion of the Spring 2022 pilot and subsequent scale reductions.

Table 11. Summary of Revised 10 Deep Dive Scales

Scale	# of items retained from Spring 2021 pilot	Estimated reliability for retained items only	# of new items added to scale for Spring 2022 pilot	Total # of items for Spring 2022 pilot
ASD	8	0.93	1	9
ASO	5	0.76	6	11
CC	9	0.91	1	10
CUS	12	0.68 (Grade 1 responses only)	4	16
HE	9	0.89	2	11
REL	10	0.90	1	11
RL	10	0.87	1	11

SCA	6	0.83	8	14
WCF	1	NA	10	11
AAL	15	NA	2	17

The Spring 2022 pilot will provide another opportunity to continue to gather on-going evidence of scale reliability and validity for supporting their intended uses within the Leaps Student Survey System. Analytic priorities for this new pilot will include:

1. Gathering data to inform item reductions to finalize 8-10 item versions of each scale
2. Piloting for the Anytime, Anywhere Learning Leap scale for the first time
3. Piloting the Customization Leap scale with a large sample of 3-8 students to reproduce the analyses described thus far in this report.
4. Capturing demographic variables for students engaging with the CUS, ASD, SCA, AAL, and WCF scales so that DIF analyses can be run for those items.
5. Gathering additional data for scales that were only administered at School 2 given known administration anomalies at this site

Once the 10 Deep Dive Leaps Scales are finalized after the Spring 2022 pilot, a rating scale measurement model will be fit to each of the scales to produce a scale score metric for each Leap. The Leaps Student Survey System employs a common-item linking approach across the Leaps Pulse Check Scale and the Deep Dives so that, after calibration, results on the Leaps Pulse Check and the Deep Dive Scales can be meaningfully compared to one another and over time. This calibration step will be essential in order to best support the intended scale interpretations and uses as outlined in the Leaps Student Survey System theory of action.

Leaps Pulse Check Scale

Scale Development

During the development of the 10 Deep Dive Leaps scales, the Transcend team recognized that it could be very valuable if a school had access to a single survey form that could support diagnostic interpretations at the school-level about learning experiences on all 10 Leaps--a Leaps Pulse Check Scale. The Pulse Check Scale would need to (a) align to our theory of action, (b) be sufficiently reliable to support class-level analyses related to the Leaps, and (c) contain enough items to support a separate interpretation for each Leap. Additionally, the Pulse Check Scale would ideally take students a fairly short amount of time to complete.

The Pulse Check Scale contains items from all of the 10 Leaps scales–AAL, ASD, ASO, CC, CUS, HE, REL, RL, SCA, and WCF. The initial design of the Pulse Check scale required that for every Leap, enough items were included to achieve at least a coefficient α of 0.70. The Pulse Check Scale, piloted in Fall 2021 was the result of making informed decisions about item selection and limited item generation in a way that balanced the known psychometric properties of the items (e.g., response distributions, item-total correlations, estimated subscale reliably), as well as equally-important content-based decisions related to construct representation.

Identifying Items for the Leaps Pulse Check Scale

To construct the Pulse Check Scale, our intention was to identify the minimum set of items to both represent the original intended meaning of each scale and produce coefficient α of at least 0.70. Table 12 outlines the items that comprised the initial proposed Pulse Check scale. Note that the new suggested item for the REL is phrased very similarly to REL_2_2, “*** the things I’m learning are important to me.” For this scale, reliability analyses were completed using the two existing items as well as using the two items plus REL_2_2.

Table 12. Shortened Leaps scales

Leap	Piloted item(s)	New item(s)
ASD	ASD_Q_1_5, ASD_Q_1_10	In school, I feel in charge of my experience. I have goals for my learning, and I have lots of choices in how I pursue those goals.

ASO	ASO_Q_1_1	*** helps me feel good about who I am. I can be myself ***.
CUS	CUS_Q_1_3, CUS_Q_1_6, CUS_Q_1_7, CUS_Q_2_5	–
HE	HE_Q_1_3, HE_Q_2_6	In school, it feels like I'm expected – and supported – to learn a ton.
RL	RL_Q_1_1, RL_Q_1_3, RL_Q_2_5	–
WCF	WCF_Q_1_1, WCF_Q_1_7	–
CC	CC_Q_4_1	I feel part of the community in school. There are a lot of people who know and care about me.
REL	REL_Q_2_3, REL_Q_2_8	In school, what I'm learning matters a lot to me.
SCA	SCA_Q_4_1	In school, we examine the problems in society – including racism and discrimination – and we take action to solve them.

For scales where the suggested items included at least two pre-existing items, Cronbach's α was calculated as a lower bound estimate of the reliability of the relevant subscale. For example, Cronbach's α for the shortened WCF scale was calculated from responses to WCF_Q_1_1 and WCF_Q_1_7, and can be interpreted as the lower bound to the reliability of total scores based upon just those two items. For these scales, Table 13 presents the lower-bound expected coefficient α.

Table 13. Anticipated Cronbach's α for reduced scales

Leap	Lower-bound Cronbach's α
ASD	0.83
CUS	0.64
HE	0.49
RL	0.69
WCF	0.75

REL	0.70*

Note: α for REL scale is based upon the three-item version noted above.

These coefficients are generally acceptable. Although CUS falls below 0.7, it is important to note that responses to CUS come from Grade 1 students and reliability is likely deflated for this sample. Although HE's reliability is the lowest, the Pulse Check scale does include an additional item, which is likely to improve reliability compared to this two-item version. Finally, RL's coefficient of 0.69 falls just barely below 0.70. Additionally, note that 0.70 is a somewhat arbitrary cutoff for individual-level reliability. For classroom-level inferences based upon the mean, even very low reliability at the individual level may not be an issue.

For the subscales that include only one preexisting item, we could not calculate coefficient α. Instead, to understand the extent to which item responses were representative of mean scores on the original scales, we calculated the correlation between responses to the relevant single item and total scores on the original, full scale (item-total correlations). For the ASO subscale, this correlation was only for the shortened five-item scale noted above. For the CC and SCA subscales, the correlation was between the item and the shortened version of the Leaps Deep Dive scale (in both cases, eight items). These correlations are found in Table 14.

Table 14. Item-total correlation for one-item reduced scales

Leap	Item-total correlation
ASO	0.72
CC	0.81
SCA	0.81

These fairly high item-total correlations, combined with the addition of more items, indicate that reliability of the short scales should be acceptable for each Leap. This led to a Pulse Check Scale of 29 items to be pilot tested in Fall 2021.

Field Test Items

In addition to the 29 item Pulse Check Scale items, a number of field test items were administered and tested using an a/b randomization design. Information on field test item administration is provided in Table 15.

Table 15. Information on field test items, Fall 2021

Scale	Item ID	N responses
WCF	V2_WCF_Q_11	1391
ASO	V2_ASO_Q_4	779
	V2_ASO_Q_9	699
SCA	V2_SCA_Q_1	829
	V2_SCA_Q_11	1781
	V2_SCA_Q_12	666
CUS	V2_CUS_Q_16	344
AAL	V2_AAL_Q_17	843

Pilot Analyses

Administration and data collection

Item responses were collected during Fall 2021. Participating schools came from two projects in which Transcend is engaged: one with rural schools, and one with schools in Texas. There was some overlap in these two projects. In total, 7662 students from grades 3-12 answered a version of the Leaps Pulse Check survey. Among these students, 56.0% came from rural schools in California, Colorado, Kentucky, Minnesota, Mississippi, and North Dakota. An additional 5.8% of the sample came from rural schools in Texas. The remainder of the sample came from non-rural schools in Texas. A full breakdown of the sample by project source can be found below in Table 16.

Table 16. Geographic breakdown of sample

State	Rural Project	Texas Project	Texas + Rural Projects
California	135	–	–
Colorado	110	–	–
Kentucky	1227	–	–

Minnesota	346	–	–
Mississippi	2077	–	–
North Dakota	398	–	–
Texas	–	2924	445
All	4293	2924	445
% of Total Sample	56.0%	38.2%	5.8%

The sample consisted of students in grades 3-12; as shown in Table 17, middle schoolers were most heavily represented in the sample, with earlier and later grades represented about evenly.

Table 17. Sample breakdown by student grade

Grade	N
3	447
4	517
5	549
6	990
7	1201
8	1143
9	598
10	512
11	469
12	388
Missing	848

In terms of race/ethnicity, the sample was fairly evenly split across Black, Hispanic/Latino and White, with a small but non-negligible number of students with multiple races listed, especially White + Hispanic/Latino. See Table 18 for a full breakdown.

Table 18. Breakdown of sample by race/ethnicity

Race/ethnicity	N
American Indian or Alaskan Native	19
Asian	42
Black or African American	2071
Hispanic or Latino	1756
Native Hawaiian or Pacific Islander	4
Two or More Races	632
White	2294
Not Reported	871

The sample was close to evenly split by gender; 3885 students were listed as male, 3634 were listed as female, and two were listed as both genders. Due to reporting issues, 844 students had no demographic data listed for either race/ethnicity or gender.

In the sample, a majority of students (n=5475) were listed as eligible for free- or reduced-price lunch, but 1341 students were missing data for this variable. Still, the majority of the sample was certainly FRL-eligible, typically used as an indicator of socioeconomic disadvantage. Pervasive missingness in school reporting of all remaining demographic variables precluded the inclusion for this information in our analyses.

Initial item and subscale analyses

Below, we report descriptive statistics and for each item: we report the number of responses, the mean response (out of five), the standard deviation (SD), and the item-total correlation (between responses to that item and responses to other items on its subscale). For item-total correlations, we calculated the correlation of responses to a given item with the total score on the *rest of the items on the subscale*, rather than the total including the relevant item. We only analyzed responses from students who saw every item listed. This means that for the item-total correlations only, we dropped any rows that include responses to some, but not all, relevant items on the Pulse Check scale. Also note that for subscales where most items were administered to most students, but some were only administered to a much smaller subset due to field testing or a/b testing, we have included versions of these analyses both including and excluding the partially-administered item(s).

We then turn to subscale properties. Here, we report the mean and distribution, as well as reliabilities, for each version of the subscale (including or excluding field-tested items). As in the pilot phase of testing, our goal was a minimum α value of 0.70.

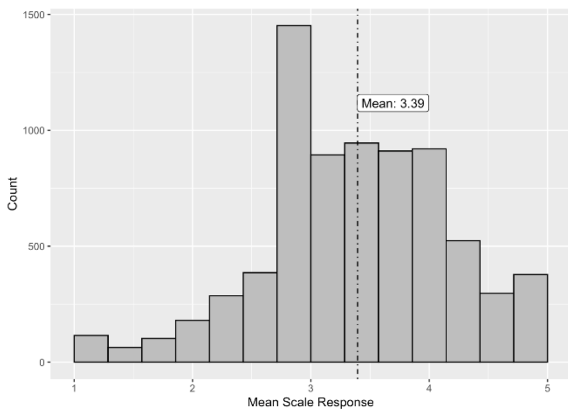
Item statistics

As shown in detail in Appendix D, the item-total correlations for the items on all subscales were quite high. A correlation magnitude of 0.30 is often used as a rule of thumb for acceptable item-total correlations, and item-total correlations for all pulse check Leaps exceed this threshold, often by quite a bit. This indicates that the subscale properties of each Leap likely support treating the individual Leap as a unidimensional scale; this was verified via exploratory factor analysis.

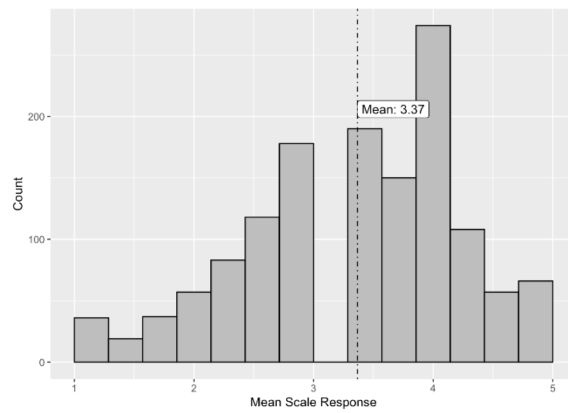
Scale means, distributions and reliabilities

Here, we briefly summarize the findings for each subscale. In most cases, mean responses appear normally distributed and are not toward the extreme ends of the subscale. The only non-normality appears to be attributable to the short length of some subscales and the fact that as a result, there are mathematically only so many possible average scores. Some distributions do have long left tails but we do not see any evidence of major non-normality such as bimodal or uniform response patterns. Figure 6 shows the distributions for each subscale, note, the shown distributions are those combinations of items that produced highest reliabilities

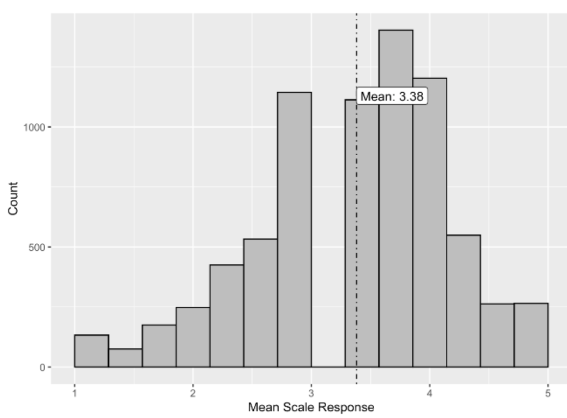
HE



WCF

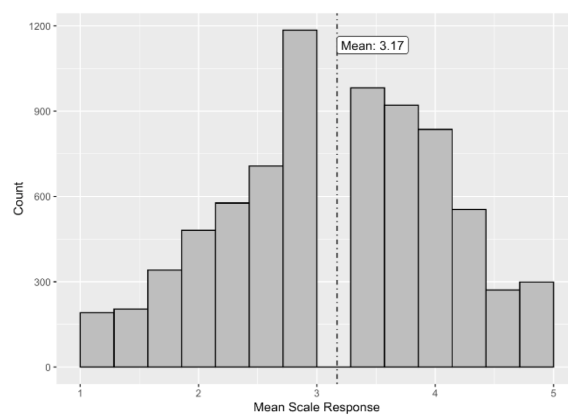


RL

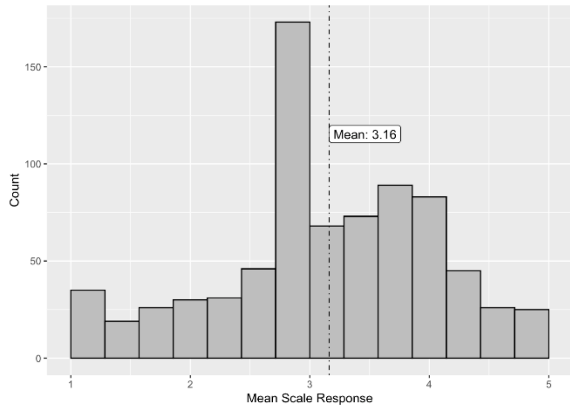


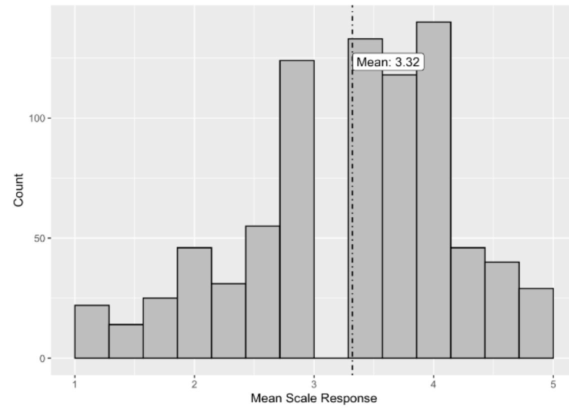
ASO

REL

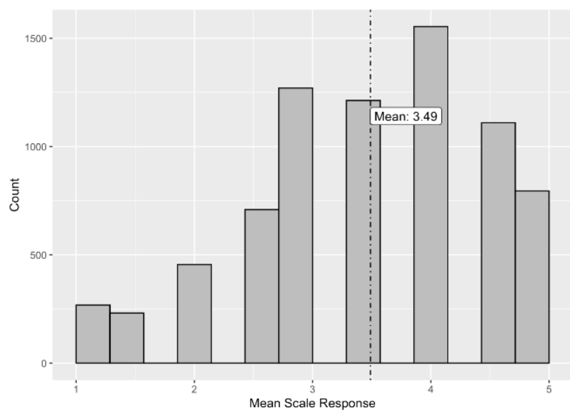


SCA

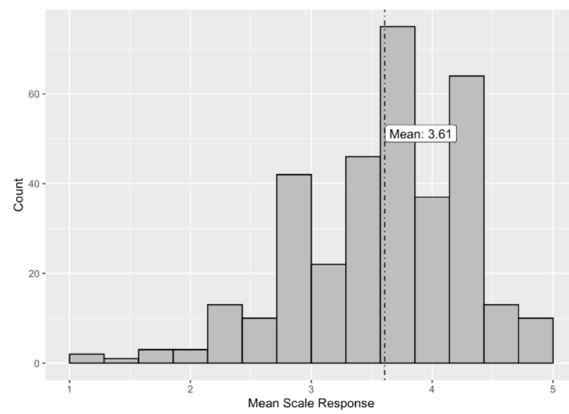




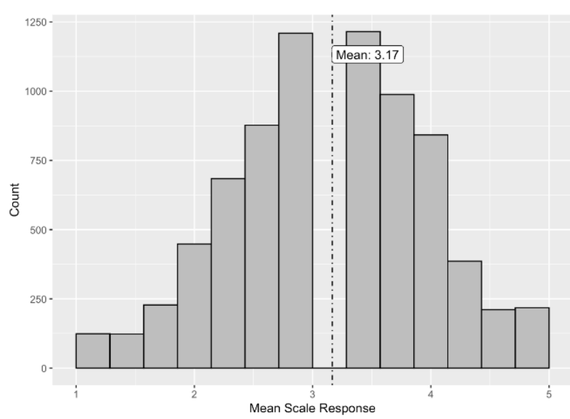
CC



CUS



ASD



AAL

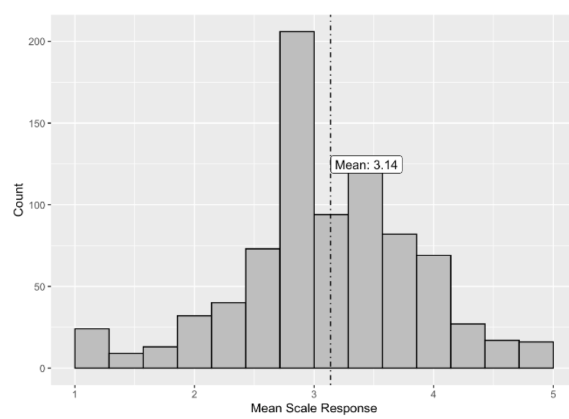


Figure 6. Mean response distributions for each Pulse Check subscale

Given the low number of items per subscale, the estimated reliability generally appears strong as shown in Table 19. When excluding the field testing items, the ASO, AAL and SCA scales produce α values below 0.7, but not by much. These lower reliabilities were incorporated into the resulting subscale recommendations. The only subscale that produced concerning results from a reliability perspective is the ASD leap, with reliability estimated well below 0.70.

Table 19. Subscale reliabilities

Subscale or subscale variation	Cronbach's α
HE	0.72
WCF w/ field test item 11	0.74
WCF w/o field test item	0.76
RL	0.72
REL	0.71
ASO w/ field test item 4	0.80
ASO w/ field test item 9	0.79
ASO w/o field test items	0.69
SCA w/ field test item 1	0.75
SCA w/ field test item 11	0.75
SCA w/ field test item 12	0.71
SCA w/o field test items	0.62
CC	0.72
CUS w/ field test item 16	0.77
CUS w/o field test item	0.73
ASD	0.56
AAL w. field test item 17	0.71
AAL w/o field test item	0.62

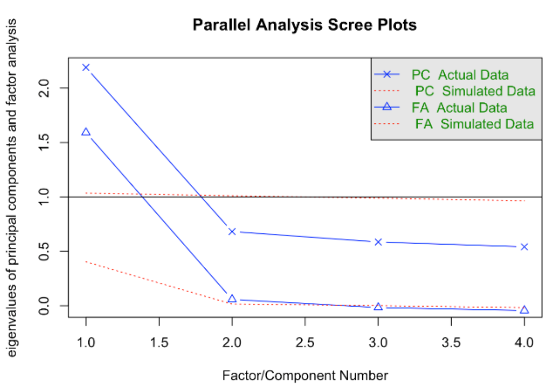
Internal structure

EFA and CFA with cross-validation

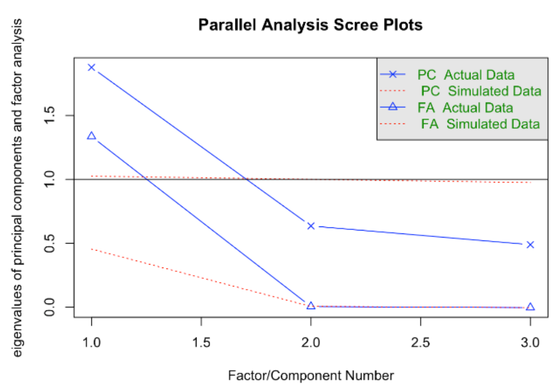
To investigate the dimensionality of each subscale, we conducted exploratory factor analyses. Exploratory factor analysis parses the variance/covariance matrix across items on a subscale to express the extent to which items appear to relate to a common underlying factor or factors. We began each analysis with a parallel analysis, which compares the eigenvalues of each factor to those found when simulating item responses at random. After parallel analysis, our next step was to separately fit unidimensional and two-dimensional factor models for each subscale. However, in most cases, the short length of the subscale led to negative degrees of freedom for a solution with two or more factors. As a result, we focused mainly on the fit of the unidimensional model. We used principal axis factoring and polychoric correlations for the analyses. For these analyses, the sample was split randomly in half, with one half used for the EFA associated with each subscale, and the other for the confirmatory factor analysis (CFA) on the entire Pulse Check scale.

Because of the inability to fit multidimensional solutions, this technical report includes only the plot from our parallel analysis and the root mean square residual (RMSR) from the unidimensional solution, shown in Figure 7 and Table 20, respectively. As in the prior section, we ran analyses for each possible version of a scale in the case where items had been field tested. Note, Figure 7 shows the resulting scree plots for the Pulse Check subscale item combinations that were ultimately recommended. Additionally, we did not include the results of analyses that involved only two items (e.g., the CC subscale), as these analyses result in a just-identified unidimensional model with no error and are not useful for understanding dimensionality.

HE

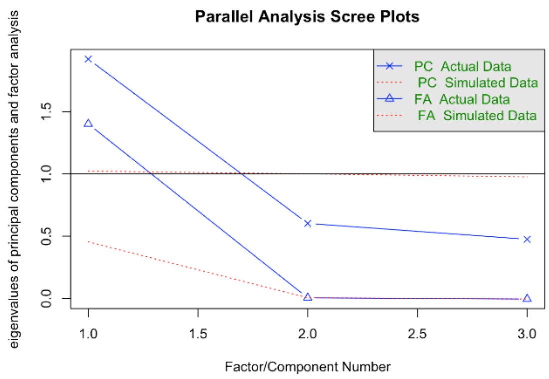


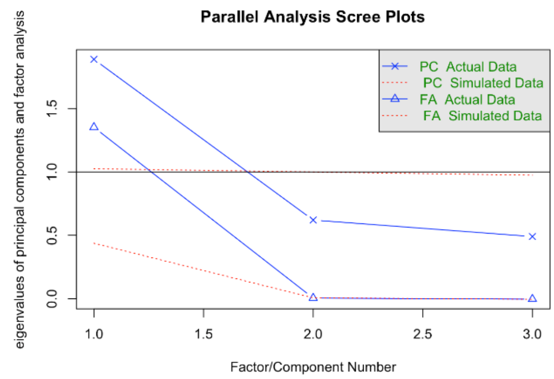
WCF



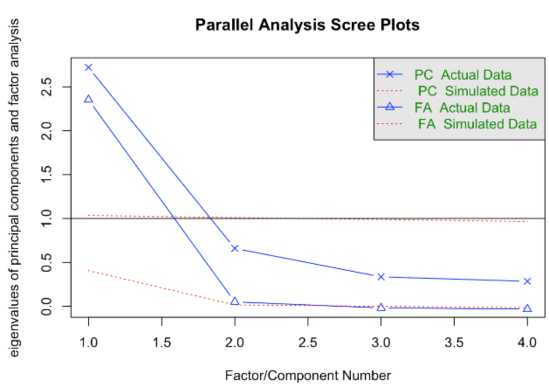
RL

REL

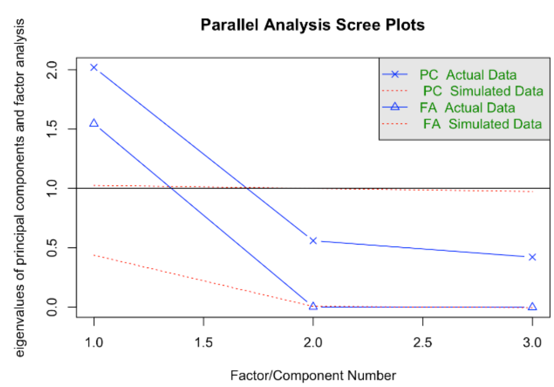

Parallel Analysis Scree Plots
PC Actual Data
PC Simulated Data
FA Actual Data
FA Simulated Data
eigenvalues of principal components and factor analysis
Factor/Component Number
0.0
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0


Parallel Analysis Scree Plots
PC Actual Data
PC Simulated Data
FA Actual Data
FA Simulated Data
eigenvalues of principal components and factor analysis
Factor/Component Number
0.0
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0

ASO


Parallel Analysis Scree Plots
PC Actual Data
PC Simulated Data
FA Actual Data
FA Simulated Data
eigenvalues of principal components and factor analysis
Factor/Component Number
0.0
0.5
1.0
1.5
2.0
2.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

SCA


Parallel Analysis Scree Plots
PC Actual Data
PC Simulated Data
FA Actual Data
FA Simulated Data
eigenvalues of principal components and factor analysis
Factor/Component Number
0.0
0.5
1.0
1.5
2.0
1.0
1.5
2.0
2.5
3.0

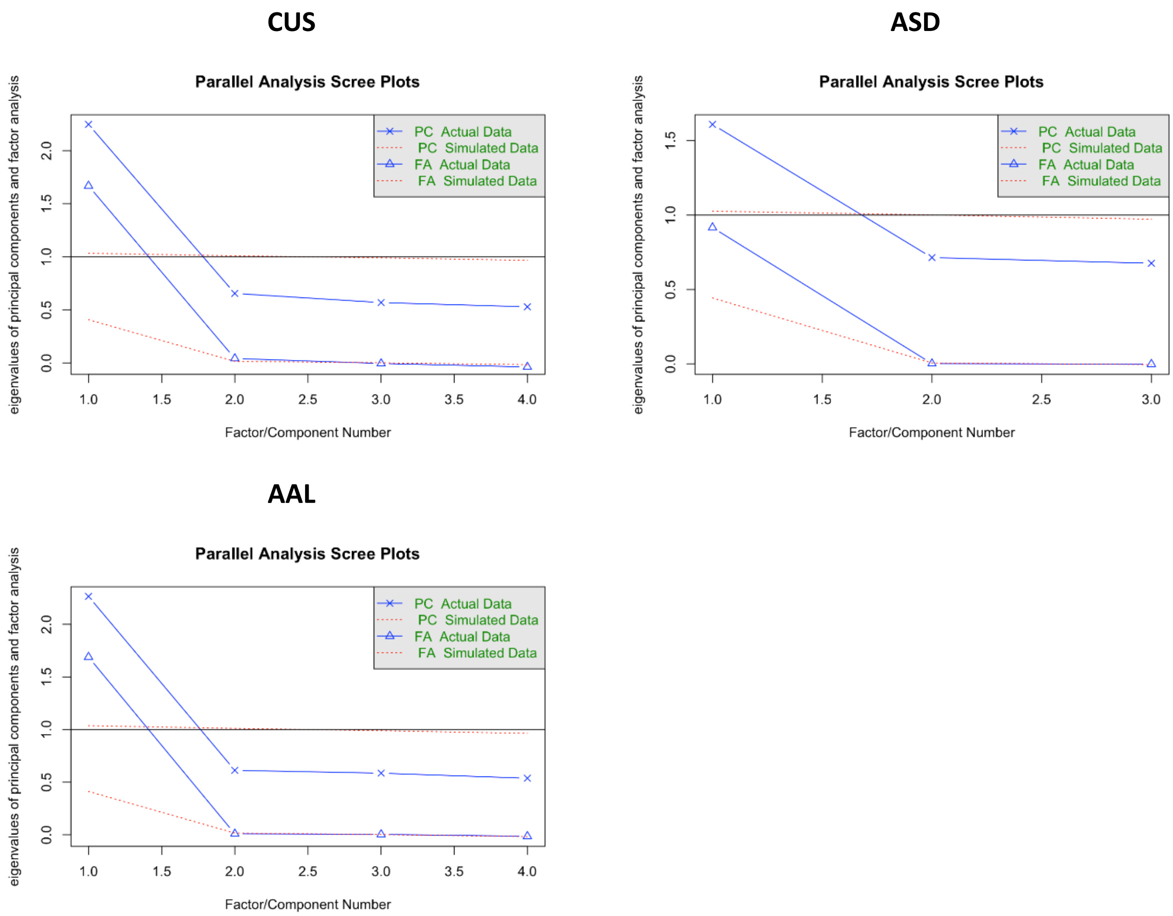


Figure 7. Scree plots from the EFA analyses on each Pulse Check subscale

Table 20. Unidimensional RMSR Values for Each Pulse Check subscale item combination

Subscale or subscale variation	RMSR
HE	0.022
WCF w/ field test item 11	0.003
WCF w/o field test item	N/A
RL	0.003
REL	0.002
ASO w/ field test item 4	0.049
ASO w/ field test item 9	0.056
ASO w/o field test items	<0.001
SCA w/ field test item 1	0.002

SCA w/ field test item 11	0.003
SCA w/ field test item 12	0.002
SCA w/o field test items	N/A
CC	N/A
CUS w/ field test item 16	0.044
CUS w/o field test item	0.071
ASD	<0.001
AAL w. field test item 17	0.057
AAL w/o field test item	0.001

We followed the EFA analyses by making a series of scale recommendations by identifying those subscales that were acceptable as-is, those subscales that required more items, and recommending which field test items should be included in the final scales. We then validated these recommendations using a CFA with the other half of the sample data to confirm the hypothesized factor structure of the entire Pulse Check scale comprising the ten underlying subscales.

Initial scale recommendations

Based on the above analyses and the reliability coefficients presented above, we established some initial scale recommendations at this point. These are listed below.

- HE scale: leave as-is.
- WCF scale: use all current items including field test item 11.
- RL scale: leave as-is.
- REL scale: leave as-is.
- ASO scale: use the version of the scale that includes field test item 4.
- SCA scale: keep all items and among the field test items, use item 11. Item 11 strikes the best balance of unidimensionality and reliability based upon the EFA and reliability coefficients above.
- CC scale: add another item. There were too few in the pulse check scale at this stage, resulting in the inability to explore dimensionality.
- CUS scale: leave as-is and do not use field test item.
- ASD scale: add another item to improve reliability. Keep all current items.

- AAL scale: use all items including field test item 17.

CFA validating findings

The next step was to conduct a CFA based upon our initial subscale recommendations. We considered two main groups of models. In all models, it was assumed that the items on each subscale load on a factor associated with that scale and do not load on any other factors. What differentiated our models was mainly the choice to include or not include a second-order factor. This factor represented the possibility that a single latent variable, tentatively referred to as "school quality," might underlie each individual Leap. We evaluated our models in terms of CFI, TLI, RMSEA and SRMR, fit statistics commonly used to evaluate CFAs. Note also that because its reliability was so low, the ASD scale was not included in the CFA. We did not think it was sensible to include since above, we recommended adding another item to the scale.

We began with a model with one second-order factor and a first-order factor for each Leap, with no correlations among the Leaps. This model's fit was not satisfactory due to residual correlations among the Leaps. This led us to consider a number of alternatives that all included a second-order factor. However, we ultimately elected to take a step back and conduct a CFA with no second-order factor, and with correlations among all Leaps. This model corresponded to the hypothesis that there is no higher-order "school quality" factor, but rather that the Leaps are distinct, yet correlated. We found this model superior to any second-order factor model we considered. CFI and TLI were equal to 0.96, RMSEA was 0.03, and SRMR was 0.03. Due to a combination of acceptable fit and parsimony, we selected this as our final model over the second-order models. According to this model, we found the following correlations among the Leaps:

Table 21. Correlations between Leaps latent variables

	HE	WCF	RL	REL	ASO	SCA	CC	CUS	AAL
HE	1	–	–	–	–	–	–	–	–
WCF	0.82	1	–	–	–	–	–	–	–
RL	0.88	0.75	1	–	–	–	–	–	–
REL	0.85	0.77	0.85	1	–	–	–	–	–
ASO	0.82	0.78	0.74	0.73	1	–	–	–	–
SCA	0.70	0.65	0.68	0.65	0.56	1	–	–	–
CC	0.84	0.74	0.70	0.71	0.88	0.56	1	–	–
CUS	0.89	0.70	0.88	0.79	0.73	0.61	0.70	1	–
AAL	0.82	0.68	0.74	0.77	0.63	0.64	0.62	0.70	1

We can see that the correlations among these variables are quite strong, but that some correlations are stronger than others. For example, High Expectations are most strongly correlated with Rigorous Learning and Customization, which is in line with the intended meaning of those scales. Similarly, the Connection and Community scale is very strongly correlated with the Affirmation of Self and Others scale–again in keeping with those scales' intended meaning.

Overall, we found that the scales appear to represent separate but correlated constructs. Because the latent variables for the Leaps were so strongly correlated, we did consider the possibility of creating a summary score based upon an average across all Leaps. The idea here was to create a composite score–not to represent a real underlying characteristic of a school, but to serve as a single metric to quickly track progress over time for schools interested in improving their Leaps outcomes across the board. One way to do this would be to conduct a Principal Components Analysis (PCA) of respondents' factor scores on each Leap. We did this first and found that a single component could explain 87% of the variance across all Leap factor scores. However, this seemed likely to introduce confusion to end users trying to make sense of the composite score without access to the factor scores underlying the PCA. A far simpler and perhaps more intuitive approach is to take each respondent's average response to each Leap on its original 1-5 scale and treat that as their score. For each respondent, we then found the mean of these Leap scores and treated that as our composite. We found a correlation of 0.97 and as shown below in Figure 8, very little indication of bias when using the average Leap score in place of the more defensible, but more complex, PCA score. Here, the blue line is a linear regression line of best fit, while the red line is a nonparametric fit curve. We can see that they are nearly identical.

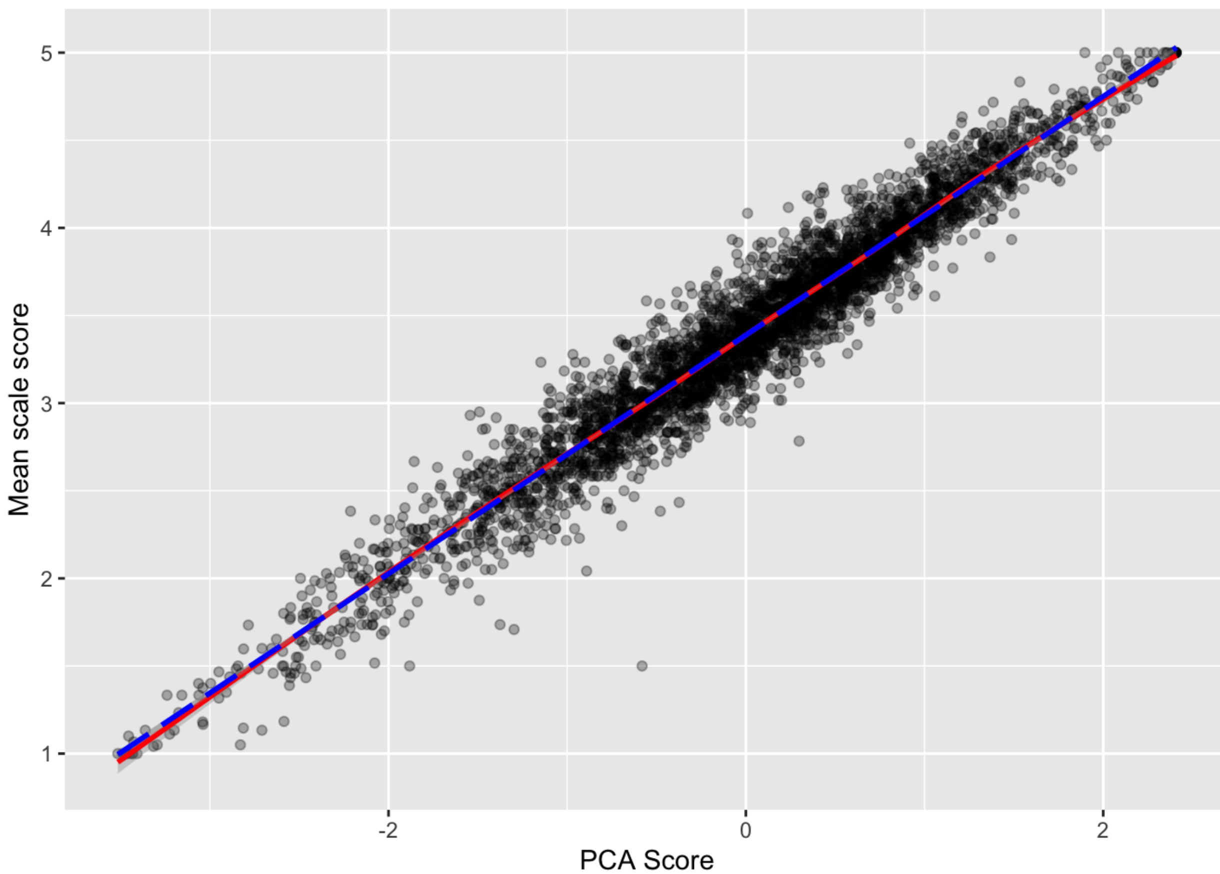


Figure 8. Correlation of mean scale score with PCA component score.

EFA of entire Pulse Check Scale

This analysis is intended to explore the extent to which individual scales emerge when all of the Leaps items are analyzed as a single dataset. The design of the Leaps Pulse Check Scale is based on the idea that each scale measures something distinct, and we have good evidence above that given this assumption, the scales behave as one would hope. However, the above-described analyses do not provide any evidence of the extent to which the scales can be distinguished based purely on students' responses on different scales.

Above, for several scales, we recommended that one of the field test items be included on the final scale. Because of the field test design, it was not possible to compute correlations among several field test items. As such, we excluded all field test items for this highly exploratory exercise. In line with the recommendation of Goretzko (2021), we used pairwise complete observations to generate the polychoric correlation matrix used in this EFA. We used promax (oblique) rotation, allowing for emergent factors to correlate. We chose this because we found the Leaps to be highly correlated in the prior section.

As above, we began with a parallel analysis, shown in Figure 9.

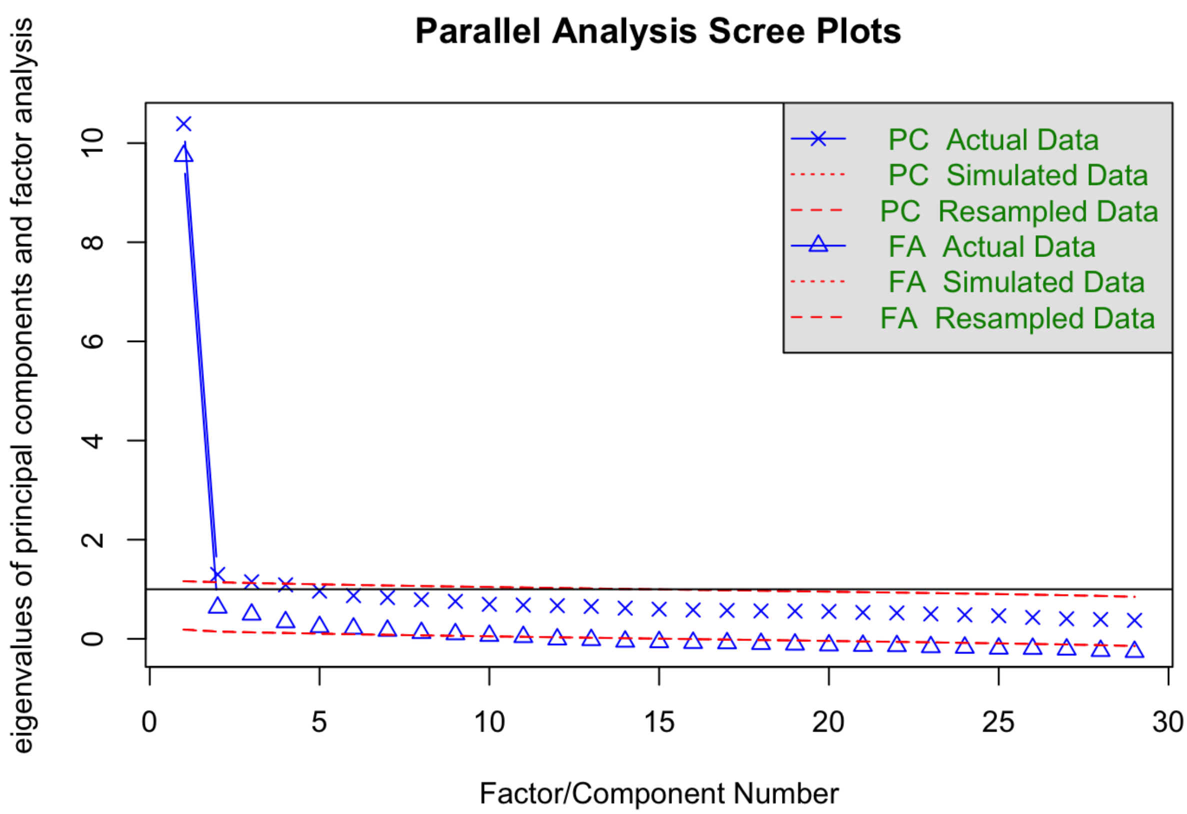


Figure 9. PCA of full Pulse Check scale.

This parallel analysis suggested a ten-factor solution, but a three-component solution for the PCA parallel analysis. The distinction between components and factors is that components attempt to account for both covariance among items and items' unique variance at the same time, while factors only attempt to account for covariance. This is very much in line with the finding above that the scales appear to be statistically distinct per our CFA, but that one can summarize across them pretty efficiently using a single composite score. From here, we fit every solution from one to eleven factors. Eleven factors are sufficient to include a strong first factor but ten separate Leap-specific factors if such a structure were to emerge.

Appendix E contains the factor loadings, fit information, and variance-accounted-for information for every solution from one to eleven factors. Here, we briefly discuss the results of the ten-factor solution, which would correspond to a separate factor for each Leap if such a structure were to emerge. We did not find that precisely, but as shown below in Table 22, several Leaps did emerge within the factor structure. This indicates that these Leaps' items were correlated with one another above-and-beyond their correlations with items on other Leaps and can be taken as evidence that the Leaps can be treated separately. Again, the factors that capture multiple Leaps do appear largely compatible with the intended meaning of the Leaps–for example, one factor includes both ASO and CC items. Although the emergent factor structure from this analysis is not perfectly aligned to the Leaps, we can say that no convincing alternative approach to Leaps scoring resulted from this analysis.

Table 22. Results of ten-factor EFA of all Pulse Check items

	PA1	PA9	PA10	PA7	PA6	PA4	PA2	PA5	PA8	PA3
V2_HE_Q_1	0.51									
V2_HE_Q_2			0.2					0.25		
V2_HE_Q_3	0.26									
V2_HE_Q_13	0.27		0.41							
V2_WCF_Q_1				0.84						
V2_WCF_Q_2				0.72						
V2_RL_Q_1	0.36							0.42		
V2_RL_Q_2								0.5	0.3	
V2_RL_Q_7								0.7		
V2_REL_Q_1					0.37					

V2_REL_Q_2					0.73					
V2_REL_Q_3	0.37				0.67					
V2_ASO_Q_1									0.26	
V2_ASO_Q_2							0.71			
V2_ASO_Q_3							0.68			
V2_SCA_Q_16						0.85				
V2_SCA_Q_17						0.6				
V2_CC_Q_1			0.62				0.2			
V2_CC_Q_2			0.71				0.27			
V2_CUS_Q_1	0.2									0.53
V2_CUS_Q_2	0.57									0.21
V2_CUS_Q_8	0.65									0.23
V2_CUS_Q_13	0.64									0.31
V2_ASD_Q_1	0.46								0.23	
V2_ASD_Q_2									0.6	
V2_ASD_Q_3									0.43	0.25
V2_AAL_Q_1		0.62								
V2_AAL_Q_2		0.72								
V2_AAL_Q_15	0.25	0.62								

Relationships to other variables

Correlations with general school attitude items

The Leaps Pulse Check Scale administration also included two items intended to capture students' attitudes toward school in general. These items and identifiers were:

- V2_GEN_1: Overall, most of the time, I love my school.
- V2_GEN_2: Overall, most of the time, I'm learning a lot in my school.

We expected that every subscale would be fairly strongly associated with either of these variables, but that Leaps related to learning directly would relate more to V2_GEN_2, while more affective Leaps would relate more to V2_GEN_1. To test this, we calculated a mean score for each subscale (based on the subscale recommendations indicated above), and then calculated the correlation between that score and each of the "general" Leap items. Table 23 summarizes these findings.

Table 23. Correlations of Pulse Check subscales with love of school and learning in school

Scale	*r* with V2_GEN_1	*r* with V2_GEN_2
HE	0.54	0.57
WCF	0.48	0.40
RL	0.46	0.53
REL	0.55	0.56
ASO	0.53	0.50
SCA	0.41	0.44
CC	0.53	0.47
CUS	0.46	0.56
ASD	0.40	0.43
AAL	0.53	0.55

Correlations tended to be similar across both general items, but we do see some evidence of stronger associations with V2_GEN_2, the general item dealing with learning, for CUS and RL, both of which do pertain to learning. Similarly, the scales less related to learning, such as WCF, were more strongly associated with the general item asking about loving school.

Analysis of variance with demographic and school-level variables

We conducted analyses of variance (ANOVAs) for all Leaps. These were intended to provide insight into the extent to which variance in students' mean response to each Leap could be accounted for using a combination of school identifiers and student demographics. Our expectation was that the school a student attends would be a significant predictor of Leaps scores. The inclusion of demographic predictors was exploratory. It is important to note that the schools in the Pulse Check Scale pilot are somewhat homogenous, as they are all either rural schools or located in Texas. Therefore, the following ANOVAs may slightly understate the amount of variance that schools would account for in a more heterogeneous group of schools. On a similar note, the reliability of the pulse check scales ranged from 0.56 to 0.80. Per classical test theory (Lord & Novick, 1968), this means that anywhere from 20% to 44% of the variance in students' scores can be attributed to random error, and this random error is by definition uncorrelated with which school a student attends or any other student variable, including their score. This means that the amount of variance for which schools could conceivably account is less than 100% of the total variance in scores.

For each scale, we fit an ANOVA with students' school as a predictor alongside a fully-interacted set of demographics: race, gender and FRL status. In all cases, the school that a student attends accounted for a statistically significant amount of variance in scale responses. Shown below in Table 24, the school level accounted for between 2% and 11% of the variance in students' responses. Detailed results of each ANOVA can be found in Appendix F. We interpret these findings as further evidence for the validity of the intended interpretations of Leaps scores, based upon findings that reflect an anticipated relationship between the Leaps and which school a student attends. We also note that gender was found to account for significant variance on several scales, which is an area for further investigation as the Leaps are introduced to more diverse groups of schools.

Table 24. School- and student-level variance in ANOVAs with school and interacted demographic predictors.

Leap	Student-level/Residual variance	School-level variance
AAL	93.0%	5.3%
ASD	95.7%	4.0%
ASO	93.8%	2.2%
CUS	92.4%	7.3%
HE	89.3%	10.5%
RL	92.7%	7.1%
WCF	94.1%	5.7%
CC	90.8%	8.8%
REL	90.9%	8.6%

SCA	92.2%	7.0%

Note: values add to less than 100%. Remainder of variance is attributable to demographic variables.

Final Pulse Check Scale Revisions

At this point, we conducted some final, ad-hoc analyses to arrive at a final set of recommendations for the next version of the Pulse Check scale. Here, we focus on the analyses used to generate these recommendations.

Shortening scales further

First, we considered scales containing more than three items. Would it be possible to reduce them to three items while retaining a reliability estimate of at least 0.70? The HE, ASO, CUS and AAL scale versions we recommended above were four items long. For each scale, we considered removing the item with the lowest item-total correlation. We found that for HE and AAL, this resulted in a Cronbach's α value below 0.70, so we did not shorten either scale. However, for ASO, removing item V2_ASO_Q_1 resulted in a Cronbach's α value of 0.86, while removing V2_CUS_Q_1 produced a α value of 0.71.

Investigating matrix sampling

For all scales, we also looked at whether it would be possible to maintain an α value of at least 0.70 using only two items from the scale. The theory behind this was that we could matrix-sample survey items to facilitate aggregate inferences about each Leap while keeping the student-level survey form as short as possible. We did not find evidence that this approach would be viable, as only the ASO scale was found to be compatible with matrix-sampling.

Final scale recommendations

Based on all of our analyses so far, we arrived at the recommendations below. They mirror the recommendations in the "Internal structure" section almost entirely, but reflect the suggestion to shorten the CUS and ASO scales.

- HE scale: leave as-is.
- WCF scale: use all current items including field test item 11.
- RL scale: leave as-is.

- REL scale: leave as-is.
- ASO scale: use a version of the scale that includes field test item 4 and does not include item 1.
- SCA scale: keep all items and among the field test items, use item 11. Item 11 strikes the best balance of unidimensionality and reliability based upon the EFA and reliability coefficients above.
- CC scale: add another item. There were too few in the Pulse Check scale at this stage, resulting in the inability to explore dimensionality.
- CUS scale: do not use field test item and remove item 1.
- ASD scale: add another item to improve reliability. Keep all current items.
- AAL scale: use all items including field test item 17.

Leaps Scaling

The Rating Scale Model

Currently, all Leaps are scored by taking the arithmetic mean of each student's responses to the items on the Leap, coded from 1-5. This approach has both advantages and disadvantages. The main advantage is that because the Leaps themselves often describe a characteristic of students' school experience that is familiar to end users, responses to the Leaps are directly interpretable (for example, it would be immediately concerning if a school found that most students strongly disagree that their teacher accepts them for who they are). However, the primary disadvantage is that the meaning of students' scale scores (calculated as their average response to the items on that scale) depends upon the particular set of items included in the scale--precluding comparability in interpretations when items are modified over time. Most pressingly, this means that one cannot easily translate between results on a Pulse Check Leap and its Deep Dive equivalent.

To address these issues in scale interpretability, we explored the use of the Rating Scale Model (Andrich, 1978) as a second way of summarizing students' Leap responses. The Rating Scale Model is an extension of the Rasch model (Rasch, 1960) for items with more than two response categories. Under the Rating Scale Model, a score for each student and a set of threshold parameters for each response category are estimated. Crucially, these parameters are all expressed on the same scale. Each item also receives a unique location; these combine with the threshold parameters to produce a probability that a student with a given score will respond in each category. The Rating Scale Model is appropriate for surveys where all items use the same response categories, which is true of all Leaps.

The Rating Scale Model provides two advantages that are difficult or impossible to achieve using raw mean scores alone. First, assuming that the model is appropriate, the Rating Scale Model provides *parameter invariance*, meaning that students' scores should not change even if the specific items they see change (beyond estimation error). This makes it possible to directly compare students' responses on a Pulse Check Leap to their subsequent responses on a Deep Dive scale, in line with the Leaps Student Survey System theory of action.

Second, for each Leap, the Rating Scale Model provides a test information function. This expresses the Fisher Information provided at various locations along the score scale. This is helpful for identifying locations on the scale that any items added to the Leap should target. As an example, Figure 10 shows the test information function for the WCF scale using only Pulse Check items. This figure shows that statistical information is highest at the middle of the scale, indicating that if additional item development were to occur, those items may be developed to target the high and low ends of the scale. This would maximize the reliability of scores across the full range of student responses on this scale.

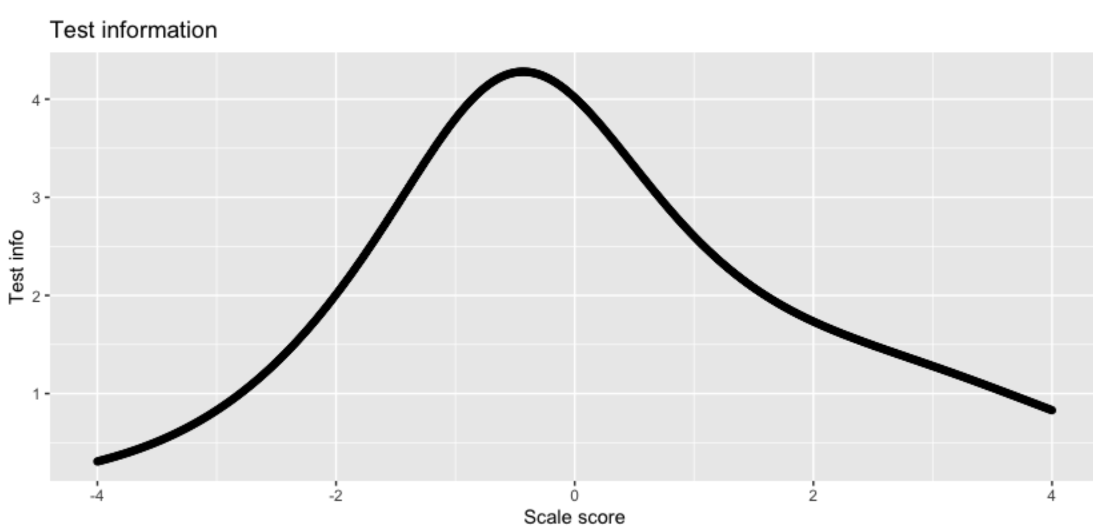


Figure 10. Example test information function for WCF pulse check scale.

However, the Rating Scale Model is not intended as a wholesale replacement for the current Leap scoring approach. As one can see on the X axis above, the scale on which the Rating Scale Model estimates student scores is not readily interpretable, especially compared to the meaningful, often actionable interpretations of raw Leaps items. Still, it is straightforward to benchmark the meaning of the score scale against the average responses of a targeted population, allowing useful norm-referenced interpretations of scores that are not possible from mean raw responses. Because the Rating Scale Model (a) allows direct comparison of Pulse Checks and Deep Dives, (b) helps guide item development, and (c) facilitates norm-referenced interpretations, we recommended it as a useful complementary scoring approach.

As of the writing of this report, we have conducted an initial exploration of the Rating Scale Model for use with the individual Leaps as operationalized within the Leaps Pulse Check Scale. We conducted this analysis using the same Pulse Check response dataset that we used for the analyses of the Pulse Checks above, and cross-validated our findings with 4816 responses gathered from students from the Rural and Texas projects in the months following the collection period for our main dataset.

For each Leap, we began by calibrating the Rating Scale Model using the *mirt* software package (Chalmers, 2012). In all cases, the model converged in fewer than 100 iterations.

High-level findings

In general, we considered these initial analyses to be a promising proof of concept for using the Rating Scale Model as part of Leaps reporting. Model fit analyses tended to indicate moderate over-fit to the model, which can be a symptom of insufficient responses at the extreme ends of the scale. However, it is important to remember that we were calibrating the models with just 2-4 items per Leap because we were using Pulse Check responses. This is reflected in our recommendations below. We also found substantial variation in the test information curves of the different Leaps, demonstrating the potential usefulness of this approach for using this information to inform future item development.

Parameter stability across calibrations

To make sure that parameter invariance held, we also re-estimated item parameters using data from a separate sample of students. For all Leaps where the items were the same in both administrations, item parameters were very similar across the two models as expected. However, the field test item(s) for two scales, SCA and ASO, were different in this second sample. For SCA, we found that the item parameters (and student-side score distribution) were still stable across calibrations. However, this was not the case for ASO; item parameters changed substantially across calibrations, as did the variance of the student-side distribution. We again take this into account in our recommendations.

Scaling recommendations

Given our initial exploration of the Rating Scale Model for the Leaps, we maintain our initial support for the usefulness of incorporating this scoring approach into reporting. We make two main recommendations.

First, we recommend repeating the scale calibration exercise described in brief above using Deep Dive data. Our exploration of parameter stability showed that with so few items on the Pulse Checks, changing even one item can substantially alter the calibrated model. Even though the student side of the calibration did not appear to affect item parameters, the item side did, indicating a violation of parameter invariance. It is likely that this could be mitigated by calibrating with the Deep Dive version of each Leap; these should contain enough items to avoid the issue we found with ASO. We recommend a thorough analysis of parameter invariance of models calibrated using the Deep Dives. The parameters from these models can then be used to link and produce a score from Pulse Checks that is directly comparable to students' Rating Scale Model-based scores on the Deep Dives. This should prove useful for tracking progress over time in schools attempting to use interventions to improve specific aspects of their students' experiences.

Second, we recommend expanding the student population beyond the Rural and Texas projects, if possible. This recommendation would be beneficial for all of the validation analyses described in this report, not just the Rating Scale Model. For the Rating Scale Model, a broader population enables more robust analyses of parameter invariance, which are warranted given our findings when calibrating the ASO scale using different field test items.

Planned Next Steps

As summarized in the executive summary, this technical report provides strong validity evidence to support the intended interpretations and uses of the Leaps Pulse Check Scale as defined within the Leaps Student Survey System Theory of Action (see Figure 1). There remains outstanding analyses and additional planned next steps to ensure that all components of the assessment system meet technical quality requirements and that the system itself is able to function as intended. The following sections outline the four additional planned areas of work: 1) finalizing the 10 Leaps Deep Dive Scales, 2) limited additional analytic support for the Leaps Pulse Check Scale, 3) investing in assessment system improvements to ensure it is functioning as intended, and 4) sharing the Deep Dives and Pulse Check Scales publicly.

Finalizing the 10 Leaps Deep Dive Scales

As noted previously, additional item development and revisions occurred prior to the current, Spring 2022 administrations of the 10 Leaps Deep Dive Scales. The analytic priorities for a final round of item and scale analyses are detailed in the section of this report entitled "Revised Leaps Deep Dive Scales." Once these analyses have been completed and the scales finalized, we plan that the Leaps Deep Dive Scales be made available in a Spanish language version.

Also, as discussed in depth in the section entitled "Leaps Scaling," the final version of the Leaps Deep Dive Scales will be calibrated using the Rasch Rating Scale Model. The three primary benefits of this score reporting approach are:

1. Supports a direct comparison of student responses to the Leaps Pulse Check Scale and subsequent administration of the Leads Deep Dives as is indicated by our Leaps Student Survey System Theory of Action;
2. Allows for on-going technical maintenance of scales and replacement of items without sacrificing comparability in interpretations across scale iterations with the added benefit of informing targeted item developing using Fisher information curves; and
3. Facilitates norm-referenced interpretations of student responses to the Leaps scales.

Once the 10 Leaps Deep Dive Scales are finalized and calibrated, those item parameters can be used to link to the Pulse Check Scale to provide the same three benefits for that scale.

Additional Analytic Support for the Pulse Check Scale

While the Leaps Pulse Check Scale is in a more final state than the 10 Deep Dives, a limited number of additional analyses are planned in order to report out on the final observed descriptive statistics and reliabilities associated with the final version of each subscale--as adapted by the recommendations enclosed within this report.

Additionally, once a sufficient number of students have had the opportunity to engage with the Spanish version of the Pulse Check Scale, analyses of measurement invariance across the languages will be run to evaluate the appropriateness of interpreting scores comparably.

Assessment System Improvements

In addition to the analytic processes described in the two prior steps, we also plan to engage in systemic improvements to the Leaps Student Survey System to ensure that the system is functioning as intended within the theory of action.

First, we are planning for significant effort to go into further refining and expanding our user-friendly reports that facilitate intended interpretations and uses associated with the Leaps Student Survey System Theory of Action. In order for the assessment tools to facilitate meaning making and lead to sustained changes within school environments, school leaders and their partners in re-design must have the information they need in a way that best serves their needs. Well-designed reports require upfront clarity about the intended purposes and uses of each instrument, as well as post-development focus groups with actual users to improve design and functionality.

Secondly, once the system is built out with final scales, interpretable scoring, and user-friendly reports, we plan to conduct an evaluation of the use of the assessment system within schools. The purpose of this is two-fold:

1. To provide insight into potential system improvements for enhancing the user experience and better contributing to the ultimate goal of the Leaps Student Survey System of transforming learner experiences, and
2. Documenting user case studies showcasing how the system can be used to drive meaningful school improvement to share with funders and other important stakeholders.

Sharing Scales and Technical Evidence Publicly

The final planned stage of this work is to prepare to publicly share the results of the Leaps Student Survey System development efforts. We believe the work that has gone into the development and validation of the Pulse Check Scale and the 10 Leaps Deep Dive scales has resulted in high-quality survey tools that have the potential to be broadly useful in many research and practical settings. As a national non-profit, part of Transcend's mission is to develop and disseminate resources that support schools in more equitably serving all students. We believe sharing the survey tools themselves has the potential to advance the mission of Transcend greatly. Below, we describe two pathways for dissemination.

First, and more immediately, once the above analyses have been completed, the final Leaps Student Survey System instruments will be published on the Transcend and Lyons Assessment Consulting websites for immediate use. At this point, it would be appropriate for both the English and Spanish versions to be shared as well as any administration manuals that offer insight to users about how the tools are intended to be administered, interpreted, and used to support school improvement efforts.

A second venue for dissemination will be the preparation and submission of the scales and their supporting validity evidence to a peer reviewed journal. The Leaps Student Survey System has benefited from the existing scientific literature related to measuring student learning experiences in school settings. We believe that the work completed as part of the development of the Leaps Student Survey System represents a significant contribution to that literature and will be shared for the purposes of advancing the science associated with the measurement of learner experiences.

References

Aldridge, J. M., & Fraser, B. J. (2008). Technology-Rich Outcomes-Focused Learning Environment Inventory (TROFLEI). In *Outcomes-Focused Learning Environments* (pp. 95-100). Brill Sense.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of school psychology*, *44*(5), 427-445.

Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British journal of educational psychology*, *72*(2), 261-278.

Burns, M. K., Vance, D., Szadokierski, I., & Stockwell, C. (2006). Student needs survey: A psychometrically sound measure of the five basic needs. *International Journal of Reality Therapy*, *25*(2), 4.

Chai, C. S., Deng, F., Tsai, P. S., Koh, J. H. L., & Tsai, C. C. (2015). Assessing multidimensional students' perceptions of twenty-first-century learning practices. *Asia Pacific Education Review*, *16*(3), 389-398.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Ferguson, R. (2010, October 14). Student perceptions of teaching effectiveness. Discussion brief from the National Center for Teacher Effectiveness and the Achievement Gap Initiative, Harvard University, Cambridge, MA.

Frymier, A. B., & Shulman, G. M. (1995). “What's in it for me?”: Increasing content relevance to enhance students' motivation. *Communication Education*, *44*(1), 40-50.

Goretzko, D. (2021). Factor retention in exploratory factor analysis with missing data. *Educational and Psychological Measurement*, 001316442110220. https://doi.org/10.1177/00131644211022031

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Massachusetts Consortium for Innovative Education Assessment. (2021). School Quality Measures: Student and Teacher Survey Scales. Retrieved from: https://www.mciea.org/school-quality-measures.html

MCIEA see Massachusetts Consortium for Innovative Education Assessment

Panorama. (2021). Panorama Social-Emotional Learning Survey. Retrived from: https://www.panoramaed.com/social-emotional-learning-sel

Pyne, J., Rozek, C. S., & Borman, G. D. (2018). Malleable Social-Psychological Academic Attitudes Survey [Database record]. Retrieved from PsycTESTS. doi: https://dx.doi.org/10.1037/t76285-000

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

United States Department of Education. (2017). ED School Climate Surveys (EDSCLS). Retrieved from: https://safesupportivelearning.ed.gov/edscls/administration

Young, M. R., Caudill, E. M., & Murphy, J. W. (2008). Evaluating experiential learning activities. *Journal for Advancement of Marketing Education*, *13*(Winter), 28-40.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendices

Appendix A: Descriptive results for Leaps Deep Dive Scales

ASD item-level statistics

Item	N	Mean	SD	Item-Total Correlation
ASD_Q_1_1	62	3.03	1.13	0.68
ASD_Q_1_2	62	3.71	1.14	0.68
ASD_Q_1_3	62	2.63	1.12	0.70
ASD_Q_1_4	62	3.39	1.26	0.79
ASD_Q_1_5	62	3.50	1.29	0.78
ASD_Q_1_6	62	3.53	1.14	0.63
ASD_Q_1_7	62	3.02	1.21	0.62
ASD_Q_1_8	62	3.77	1.14	0.67
ASD_Q_1_9	62	3.23	1.37	0.76
ASD_Q_1_10	62	3.45	1.35	0.81
ASD_Q_2_1	62	3.66	1.25	0.79
ASD_Q_2_2	62	3.56	1.18	0.65
ASD_Q_2_3	62	3.55	1.21	0.76
ASD_Q_2_4	62	3.66	1.19	0.72
ASD_Q_2_5	62	3.79	1.22	0.74

ASO item-level statistics

Item	N	Mean	SD	Item-Total Correlation
ASO_Q_1_1	297	3.32	1.03	0.38
ASO_Q_1_2	297	3.63	0.95	0.48
ASO_Q_1_3	297	4.23	0.91	0.49
ASO_Q_1_4	297	4.37	0.94	0.50
ASO_Q_1_5	297	4.28	0.84	0.58
ASO_Q_2_1	297	3.57	0.94	0.33
ASO_Q_2_2	297	4.14	0.80	0.63
ASO_Q_2_3	297	4.13	0.84	0.53
ASO_Q_2_4	297	3.84	1.10	0.40

ASO_Q_2_5	297	3.77	1.15	0.40
ASO_Q_3_1	297	4.21	0.91	0.54
ASO_Q_4_1	297	3.97	1.00	0.62
ASO_Q_4_2	297	4.31	0.91	0.57
ASO_Q_5_1	297	3.81	0.88	0.50
ASO_Q_5_2	297	3.75	0.94	0.44

CC item-level statistics

Item	N	Mean	SD	Item-total correlation
CC_Q_1_1	483	3.46	1.11	0.63
CC_Q_1_3	483	3.83	1.04	0.62
CC_Q_1_4	483	3.57	1.03	0.72
CC_Q_1_5	483	3.54	1.02	0.71
CC_Q_1_6	483	4.10	0.98	0.56
CC_Q_1_7	483	3.70	1.06	0.75
CC_Q_1_8	483	3.65	0.97	0.68
CC_Q_1_9	483	3.60	0.98	0.66
CC_Q_1_10	483	3.28	1.12	0.66
CC_Q_1_11	483	3.79	0.97	0.63
CC_Q_2_1	483	3.26	1.14	0.71
CC_Q_3_1	483	3.22	1.10	0.70
CC_Q_4_1	483	3.63	1.15	0.74

CUS item-level statistics

Item	N	Mean	SD	Item-total correlation
CUS_Q_1_1	34	3.15	1.02	0.61
CUS_Q_1_2	34	3.09	1.19	0.34
CUS_Q_1_3	34	3.06	1.07	0.39
CUS_Q_1_4	34	2.79	1.25	0.41
CUS_Q_1_5	34	3.03	0.97	0.41
CUS_Q_1_6	34	3.12	1.01	0.53
CUS_Q_1_7	34	3.35	1.10	0.65

CUS_Q_2_1	34	3.5	0.90	0.04
CUS_Q_2_2	34	3.29	0.91	0.11
CUS_Q_2_3	34	3.68	0.64	0.24
CUS_Q_2_4	34	3.74	0.67	-0.10
CUS_Q_2_5	34	3.53	0.83	0.51
CUS_Q_3_1	34	3.47	0.90	0.21
CUS_Q_3_2	34	3.71	0.72	0.39

HE item-level statistics

Item	N	Mean	SD	Item-total correlation
HE_Q_1_1	314	4.06	0.84	0.31
HE_Q_1_2	314	3.86	0.90	0.45
HE_Q_1_3	314	3.97	0.83	0.53
HE_Q_1_4	314	4.11	0.84	0.38
HE_Q_1_5	314	4.08	0.88	0.62
HE_Q_1_6	314	3.99	1.00	0.60
HE_Q_2_1	314	3.74	0.91	0.50
HE_Q_2_2	314	3.80	1.02	0.63
HE_Q_2_3	314	3.87	0.94	0.61
HE_Q_2_4	314	3.96	0.82	0.66
HE_Q_2_5	314	3.97	0.86	0.62
HE_Q_2_6	314	3.61	0.97	0.58
HE_Q_3_1	314	3.33	1.06	0.48
HE_Q_4_1	314	3.56	0.95	0.30

REL item-level statistics

Item	N	Mean	SD	Item-total correlation
REL_Q_1_1	112	3.21	1.20	0.70
REL_Q_1_4	112	3.04	1.23	0.69
REL_Q_1_5	112	3.44	1.15	0.68
REL_Q_2_1	112	3.00	1.20	0.74
REL_Q_2_4	112	3.54	1.18	0.61
REL_Q_2_6	112	3.02	1.26	0.56

REL_Q_1_2	112	3.63	1.23	0.70
REL_Q_1_3	112	3.52	1.14	0.57
REL_Q_2_2	112	3.60	1.13	0.55
REL_Q_2_3	112	3.19	1.14	0.64
REL_Q_2_5	112	3.63	1.07	0.56
REL_Q_2_7	112	2.96	1.19	0.62
REL_Q_2_8	112	3.10	1.22	0.65

RL item-level statistics

Item	N	Mean	SD	Item-total correlation
RL_Q_1_1	965	3.53	1.09	0.61
RL_Q_1_2	965	3.39	1.09	0.60
RL_Q_1_3	965	3.72	1.09	0.60
RL_Q_1_4	965	3.79	1.02	0.63
RL_Q_1_5	965	3.37	1.05	0.66
RL_Q_1_6	965	3.51	1.01	0.60
RL_Q_1_7	965	3.66	0.97	0.55
RL_Q_1_8	965	3.5	1.01	0.54
RL_Q_2_1	965	3.04	1.18	0.60
RL_Q_2_2	965	3.32	1.19	0.61
RL_Q_2_3	965	3.07	1.18	0.59

SCA item-level statistics

Item	N	Mean	SD	Item-total correlation
SCA_Q_1_1	62	3.76	0.95	0.71
SCA_Q_1_2	62	3.74	0.99	0.42
SCA_Q_1_3	62	3.63	1.04	0.77
SCA_Q_1_4	62	3.56	1.14	0.62
SCA_Q_1_5	62	3.95	0.98	0.75
SCA_Q_2_1	62	3.95	0.90	0.65
SCA_Q_3_1	62	3.89	1.01	0.64
SCA_Q_4_1	62	4.05	0.82	0.75

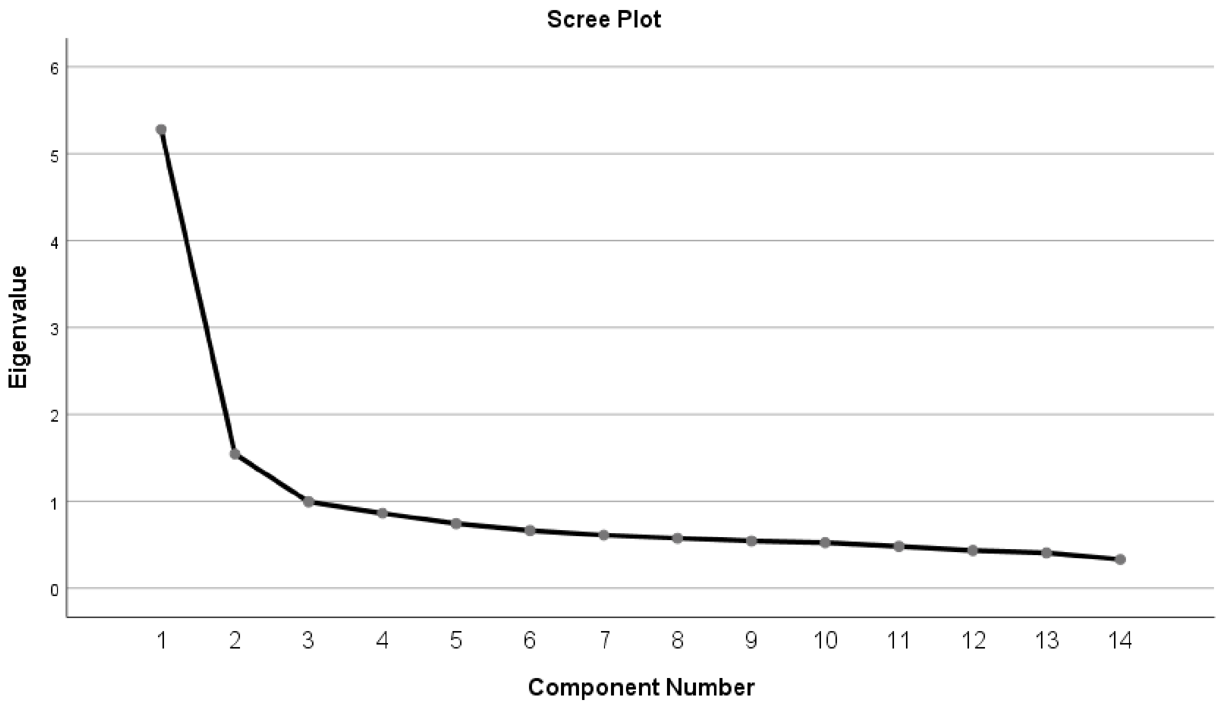
WCF item-level statistics

Item	N	Mean	SD	Item-total correlation
WCF_Q_1_1	1268	3.07	1.22	0.71
WCF_Q_1_2	1268	3.61	1.08	0.64
WCF_Q_1_3	1268	3.27	1.19	0.60
WCF_Q_1_4	1268	3.9	0.97	0.49
WCF_Q_1_5	1268	2.76	1.29	0.67
WCF_Q_1_6	1268	3.48	1.07	0.60
WCF_Q_1_7	1268	3.17	1.19	0.69
WCF_Q_1_8	1268	3.56	1.09	0.64
WCF_Q_2_1	1268	3.16	1.17	0.54

Appendix B: Leaps Deep Dives Exploratory Factor Analysis Results

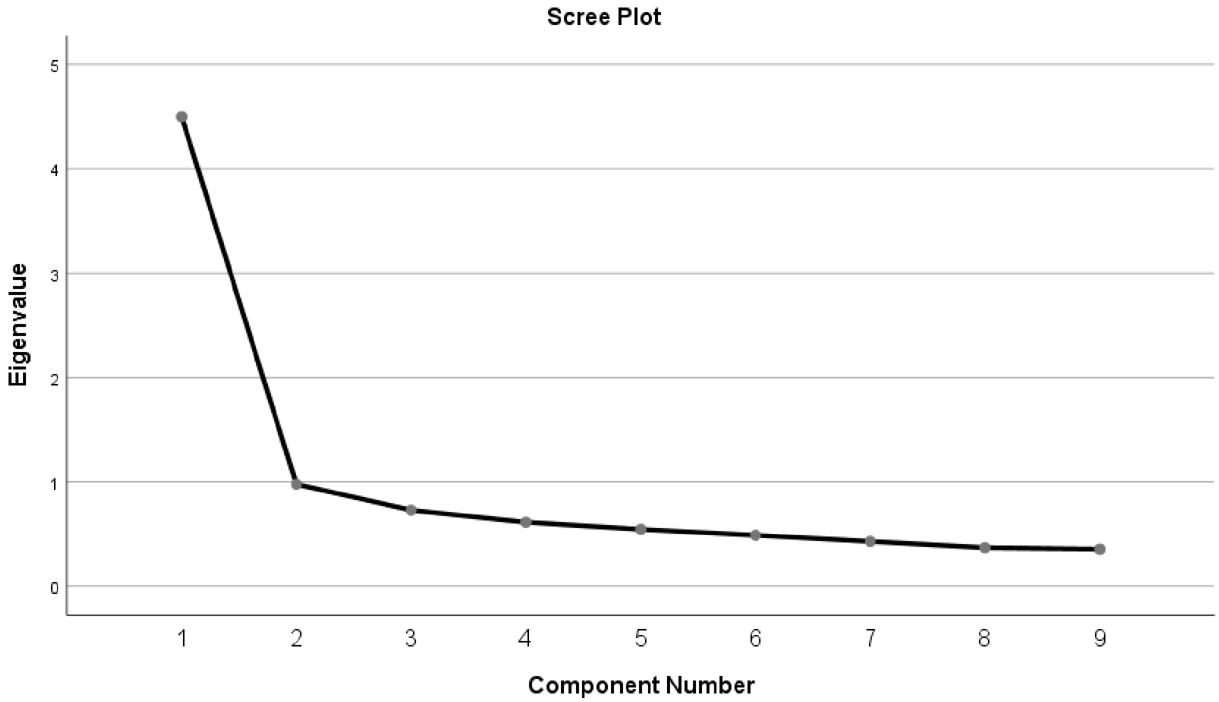
These analyses used PCA extraction.

HE scale



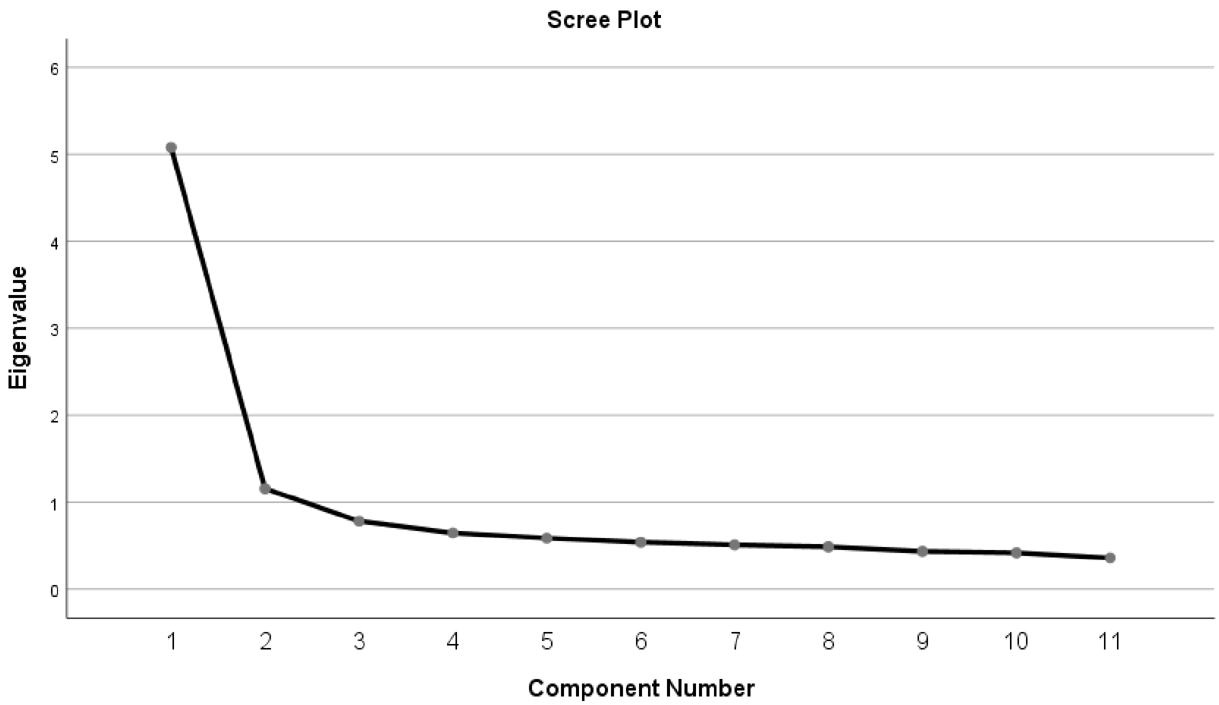
Component	Eigenvalue	% of Variance	Cumulative %
1	5.281	37.722	37.722
2	1.545	11.036	48.758
3	.994	7.103	55.860
4	.863	6.164	62.024

WCF scale



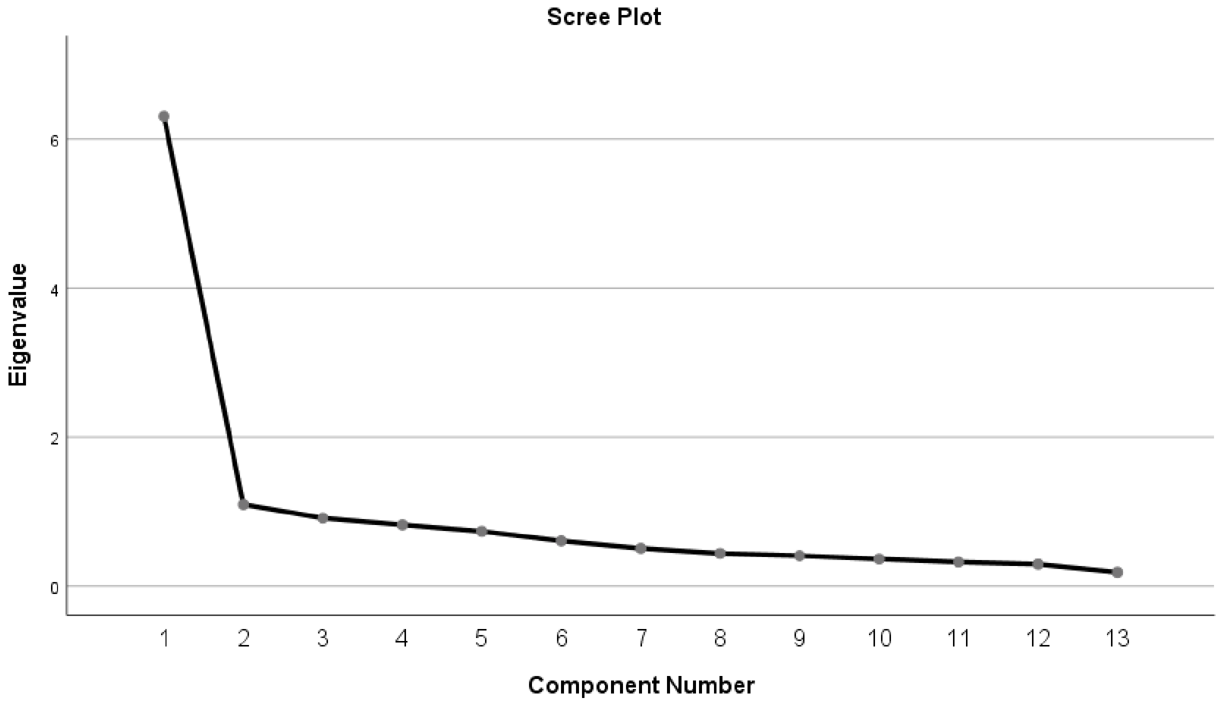
Component	Eigenvalue	% of Variance	Cumulative %
1	4.500	49.996	49.996
2	.977	10.852	60.848
3	.728	8.092	68.939
4	.612	6.804	75.743

RL scale



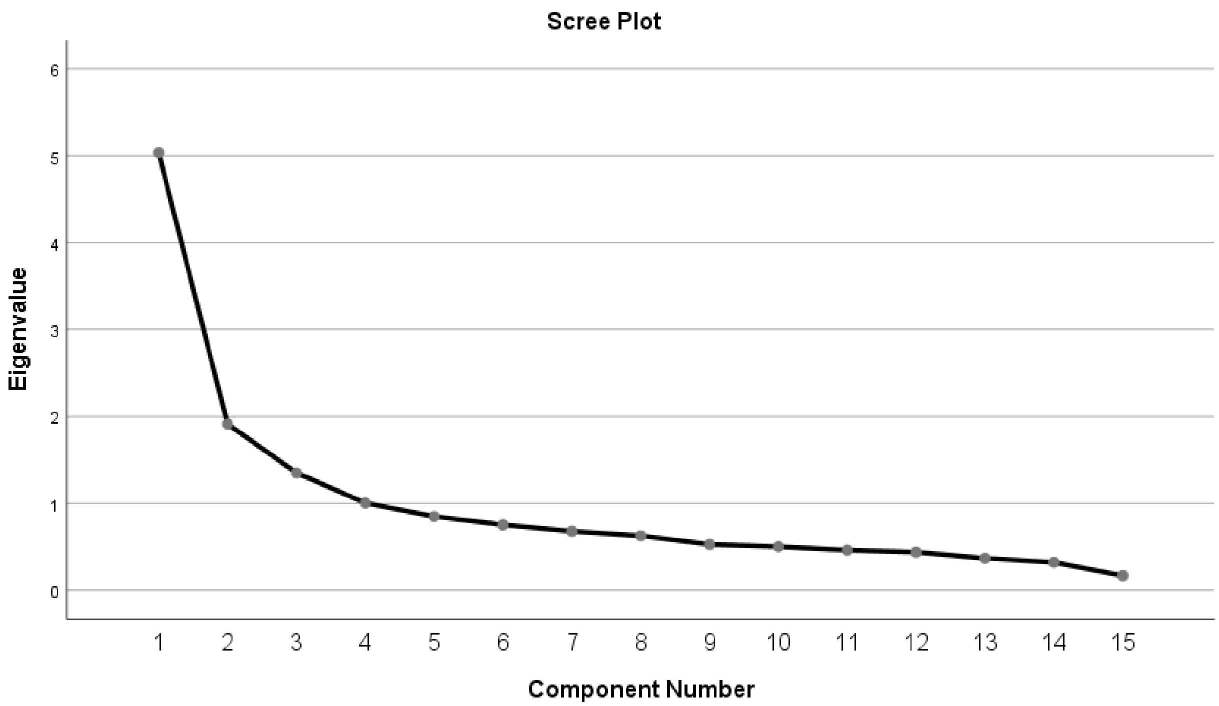
Component	Eigenvalue	% of Variance	Cumulative %
1	5.081	46.191	46.191
2	1.156	10.512	56.703
3	.782	7.112	63.815
4	.646	5.875	69.689

REL scale



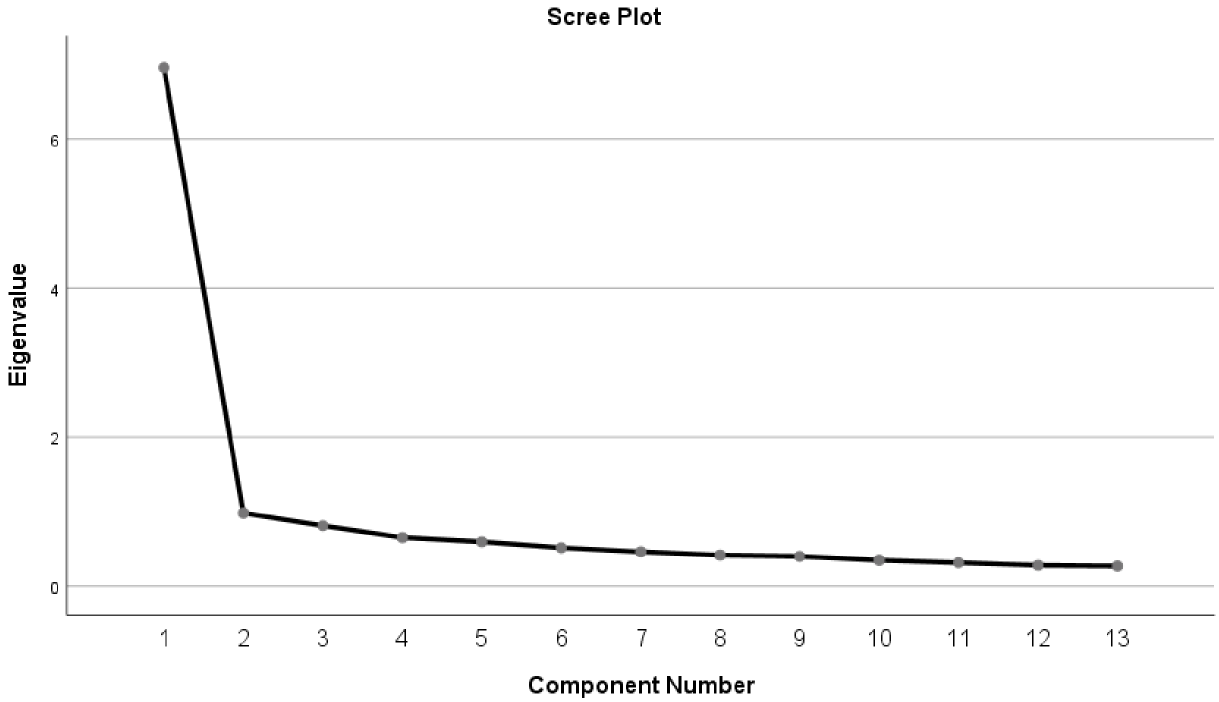
Component	Eigenvalue	% of Variance	Cumulative %
1	6.305	48.504	48.504
2	1.096	8.427	56.931
3	.913	7.023	63.953
4	.823	6.330	70.283

ASO scale



Component	Eigenvalue	% of Variance	Cumulative %
1	5.037	33.580	33.580
2	1.914	12.763	46.342
3	1.352	9.013	55.355
4	1.005	6.699	62.054

CC scale



Component	Eigenvalue	% of Variance	Cumulative %
1	6.957	53.515	53.515
2	.981	7.547	61.063
3	.812	6.244	67.306
4	.651	5.010	72.317

Appendix C: Statistical significance and magnitude of DIF for all items/scales

In tables below, * indicates significance at the $p < 0.05$ level and ** indicates significance at the $p < 0.01$ level.

HE scale

DIF by Gender

Item	Pseudo-r^2 difference	DIF effect size
HE_Q_1_1	0.005	trivial
HE_Q_1_2	0.003	trivial
HE_Q_1_3	0.023*	trivial
HE_Q_1_4	0.003	trivial
HE_Q_1_5	0.004	trivial
HE_Q_1_6	0.002	trivial
HE_Q_2_1	0.024*	trivial
HE_Q_2_2	0.011	trivial
HE_Q_2_3	0.016	trivial
HE_Q_2_4	0.006	trivial
HE_Q_2_5	0.003	trivial
HE_Q_2_6	0.006	trivial
HE_Q_3_1	0.005	trivial
HE_Q_4_1	0.003	trivial

WCF scale

DIF by Gender

Item	Pseudo-r^2 difference	DIF effect size
WCF_Q_1_1	0	trivial
WCF_Q_1_2	0	trivial
WCF_Q_1_3	0.001	trivial
WCF_Q_1_4	0.003	trivial
WCF_Q_1_5	0.002	trivial
WCF_Q_1_6	0.006**	trivial
WCF_Q_1_7	0	trivial

WCF_Q_1_8	0.007**	trivial
WCF_Q_2_1	0.005*	trivial

DIF by Race/Ethnicity: Black and White Subgroups

Item	Pseudo-r^2 difference	DIF effect size
WCF_Q_1_1	0.001	trivial
WCF_Q_1_2	0.001	trivial
WCF_Q_1_3	0	trivial
WCF_Q_1_4	0.003*	trivial
WCF_Q_1_5	0.001	trivial
WCF_Q_1_6	0.002	trivial
WCF_Q_1_7	0.003*	trivial
WCF_Q_1_8	0	trivial
WCF_Q_2_1	0.001	trivial

DIF by Race/Ethnicity: Hispanic/Latino and White Subgroups

Item	Pseudo-r^2 difference	DIF effect size
WCF_Q_1_1	0.005**	trivial
WCF_Q_1_2	0	trivial
WCF_Q_1_3	0.005**	trivial
WCF_Q_1_4	0.013**	trivial
WCF_Q_1_5	0	trivial
WCF_Q_1_6	0.01**	trivial
WCF_Q_1_7	0	trivial
WCF_Q_1_8	0.001	trivial
WCF_Q_2_1	0.004*	trivial

RL scale

DIF by Gender

Item	Pseudo-r^2 difference	DIF effect size
RL_Q_1_1	0.008**	trivial
RL_Q_1_2	0.004	trivial
RL_Q_1_3	0.003	trivial

RL_Q_1_4	0.002	trivial
RL_Q_1_5	0.002	trivial
RL_Q_1_6	0.002	trivial
RL_Q_1_7	0.004*	trivial
RL_Q_1_8	0	trivial
RL_Q_2_1	0.001	trivial
RL_Q_2_2	0	trivial
RL_Q_2_3	0	trivial

DIF by Race/Ethnicity: Black and White Subgroups

Item	Pseudo-r^2 difference	DIF effect size
RL_Q_1_1	0.001	trivial
RL_Q_1_2	0	trivial
RL_Q_1_3	0.005**	trivial
RL_Q_1_4	0.003**	trivial
RL_Q_1_5	0.001	trivial
RL_Q_1_6	0	trivial
RL_Q_1_7	0.002	trivial
RL_Q_1_8	0	trivial
RL_Q_2_1	0.002	trivial
RL_Q_2_2	0.001	trivial
RL_Q_2_3	0.001	trivial

DIF by Race/Ethnicity: Hispanic/Latino and White Subgroups

Item	Pseudo-r^2 difference	DIF effect size
RL_Q_1_1	0	trivial
RL_Q_1_2	0.001	trivial
RL_Q_1_3	0.006*	trivial
RL_Q_1_4	0.001	trivial
RL_Q_1_5	0	trivial
RL_Q_1_6	0.001	trivial
RL_Q_1_7	0	trivial

RL_Q_1_8	0.001	trivial
RL_Q_2_1	0.002	trivial
RL_Q_2_2	0	trivial
RL_Q_2_3	0	trivial

REL scale

DIF by Gender

Item	Pseudo-r^2 difference	DIF effect size
REL_Q_1_1	0.089**	small
REL_Q_1_2	0.002	trivial
REL_Q_1_3	0.008	trivial
REL_Q_1_4	0.033*	trivial
REL_Q_1_5	0.001	trivial
REL_Q_2_1	0.037*	trivial
REL_Q_2_2	0.005	trivial
REL_Q_2_3	0.03	trivial
REL_Q_2_4	0.018	trivial
REL_Q_2_5	0.007	trivial
REL_Q_2_6	0.012	trivial
REL_Q_2_7	0.006	trivial
REL_Q_2_8	0.01	trivial

DIF by Race/Ethnicity: Hispanic/Latino and Black Subgroups

Item	Pseudo-r^2 difference	DIF effect size
REL_Q_1_1	0.004	trivial
REL_Q_1_2	0.003	trivial
REL_Q_1_3	0.012	trivial
REL_Q_1_4	0.012	trivial
REL_Q_1_5	0.001	trivial
REL_Q_2_1	0.002	trivial
REL_Q_2_2	0.008	trivial
REL_Q_2_3	0.001	trivial

REL_Q_2_4	0.003	trivial
REL_Q_2_5	0.017	trivial
REL_Q_2_6	0.002	trivial
REL_Q_2_7	0.042	trivial
REL_Q_2_8	0.005	trivial

ASO scale

DIF by Gender

Item	Pseudo-r^2 difference	DIF effect size
ASO_Q_1_1	0.021*	trivial
ASO_Q_1_2	0.01	trivial
ASO_Q_1_3	0.007	trivial
ASO_Q_1_4	0.009	trivial
ASO_Q_1_5	0.007	trivial
ASO_Q_2_1	0.016*	trivial
ASO_Q_2_2	0.001	trivial
ASO_Q_2_3	0.006	trivial
ASO_Q_2_4	0.007	trivial
ASO_Q_2_5	0.004	trivial
ASO_Q_3_1	0.003	trivial
ASO_Q_4_1	0.003	trivial
ASO_Q_4_2	0.001	trivial
ASO_Q_5_1	0.006	trivial
ASO_Q_5_2	0.006	trivial

CC scale

DIF by Gender

Item	Pseudo-r^2 difference	DIF effect size
CC_Q_1_1	0.013*	trivial
CC_Q_1_2	0.001	trivial
CC_Q_1_3	0.002	trivial
CC_Q_1_4	0.003	trivial

CC_Q_1_5	0.005	trivial
CC_Q_1_6	0.003	trivial
CC_Q_1_7	0.005	trivial
CC_Q_1_8	0.001	trivial
CC_Q_1_9	0.007	trivial
CC_Q_1_10	0.005	trivial
CC_Q_1_11	0.006	trivial
CC_Q_2_1	0.006	trivial
CC_Q_3_1	0.007*	trivial
CC_Q_4_1	0	trivial

DIF by Race/Ethnicity: Hispanic/Latino and White Subgroups

Item	Pseudo-r^2 difference	DIF effect size
CC_Q_1_1	0.002	trivial
CC_Q_1_2	0.002	trivial
CC_Q_1_3	0	trivial
CC_Q_1_4	0.002	trivial
CC_Q_1_5	0	trivial
CC_Q_1_6	0.005	trivial
CC_Q_1_7	0.002	trivial
CC_Q_1_8	0.001	trivial
CC_Q_1_9	0	trivial
CC_Q_1_10	0.004	trivial
CC_Q_1_11	0.008*	trivial
CC_Q_2_1	0.003	trivial
CC_Q_3_1	0.007	trivial
CC_Q_4_1	0.004	trivial

DIF by Race/Ethnicity: Hispanic/Latino and Black Subgroups

Item	Pseudo-r^2 difference	DIF effect size
CC_Q_1_1	0.009*	trivial
CC_Q_1_2	0	trivial

CC_Q_1_3	0.003	trivial
CC_Q_1_4	0	trivial
CC_Q_1_5	0.004	trivial
CC_Q_1_6	0.005	trivial
CC_Q_1_7	0.001	trivial
CC_Q_1_8	0	trivial
CC_Q_1_9	0.004	trivial
CC_Q_1_10	0	trivial
CC_Q_1_11	0.001	trivial
CC_Q_2_1	0.007	trivial
CC_Q_3_1	0.003	trivial
CC_Q_4_1	-0.001	trivial

Appendix D: Item-total correlations for sub-scales within Pulse Check Scale

HE scale

Item	N responses	Mean response	SD	Item-total cor.
V2_HE_Q_1	7587	3.71	0.98	0.48
V2_HE_Q_2	7590	3.18	1.11	0.51
V2_HE_Q_3	7561	3.43	1.00	0.52
V2_HE_Q_13	7589	3.25	1.17	0.54

WCF scale

The WCF scale included one field test item, item 11. First, the scale including this item:

Item	N responses	Mean response	SD	Item-total cor.
V2_WCF_Q_1	7607	3.25	1.14	0.55
V2_WCF_Q_2	7603	3.28	1.13	0.56
V2_WCF_Q_11	1391	3.53	1.14	0.46

Next, without the field test item:

Item	N responses	Mean response	SD	Item-total cor.
V2_WCF_Q_1	7607	3.25	1.14	0.61
V2_WCF_Q_2	7603	3.28	1.13	0.61

RL scale

Item	N responses	Mean response	SD	Item-total cor.
V2_RL_Q_1	7604	3.53	1.01	0.51
V2_RL_Q_2	7587	3.13	1.10	0.51
V2_RL_Q_7	7594	3.49	1.00	0.59

REL scale

Item	N responses	Mean response	SD	Item-total cor.
V2_REL_Q_1	7608	2.94	1.19	0.48
V2_REL_Q_2	7605	3.09	1.18	0.57
V2_REL_Q_3	7593	3.48	1.22	0.53

ASO scale

Items 4 and 9 were a/b tested, so we have separate analyses for the version of the scale including 4 and the version including 9, as well as a version including neither.

Including item 4:

Item	N responses	Mean response	SD	Item-total cor.
V2_ASO_Q_1	7590	2.95	1.13	0.45
V2_ASO_Q_2	7599	3.60	1.11	0.68
V2_ASO_Q_3	7608	3.44	1.20	0.70
V2_ASO_Q_4	779	3.21	1.14	0.76

Including item 9:

Item	N responses	Mean response	SD	Item-total cor.
V2_ASO_Q_1	7590	2.95	1.13	0.44
V2_ASO_Q_2	7599	3.60	1.11	0.64
V2_ASO_Q_3	7608	3.44	1.20	0.58
V2_ASO_Q_9	699	3.36	1.10	0.68

Including neither field test item:

Item	N responses	Mean response	SD	Item-total cor.
V2_ASO_Q_1	7590	2.95	1.13	0.38
V2_ASO_Q_2	7599	3.60	1.11	0.57

V2_ASO_Q_3	7608	3.44	1.20	0.56

SCA scale

There were three field test items: 1, 11 and 12.

Keeping item 1:

Item	N responses	Mean response	SD	Item-total cor.
V2_SCA_Q_1	829	3.26	1.12	0.66
V2_SCA_Q_16	7266	3.35	1.11	0.45
V2_SCA_Q_17	7609	3.30	1.14	0.59

Keeping item 11:

Item	N responses	Mean response	SD	Item-total cor.
V2_SCA_Q_11	1781	3.17	1.29	0.61
V2_SCA_Q_16	7266	3.35	1.11	0.52
V2_SCA_Q_17	7609	3.30	1.14	0.62

Keeping item 12:

Item	N responses	Mean response	SD	Item-total cor.
V2_SCA_Q_12	666	2.72	1.17	0.53
V2_SCA_Q_16	7266	3.35	1.11	0.57
V2_SCA_Q_17	7609	3.30	1.14	0.57

Not using any field test items:

Item	N responses	Mean response	SD	Item-total cor.
V2_SCA_Q_16	7266	3.35	1.11	0.44
V2_SCA_Q_17	7609	3.30	1.14	0.44

CC scale

Item	N responses	Mean response	SD	Item-total cor.
V2_CC_Q_1	7619	3.51	1.08	0.56
V2_CC_Q_2	7616	3.47	1.23	0.56

CUS scale

There was one field test item, item 16. First, statistics including the field test item are included below.

Item	N responses	Mean response	SD	Item-total cor.
V2_CUS_Q_1	7590	3.12	1.16	0.53
V2_CUS_Q_2	7568	3.65	0.91	0.56
V2_CUS_Q_8	7577	3.75	1.04	0.52
V2_CUS_Q_13	7586	3.70	0.98	0.59
V2_CUS_Q_16	344	3.30	1.01	0.61

Next, the scale without the field test item:

Item	N responses	Mean response	SD	Item-total cor.
V2_CUS_Q_1	7590	3.12	1.16	0.48
V2_CUS_Q_2	7568	3.65	0.91	0.56
V2_CUS_Q_8	7577	3.75	1.04	0.53
V2_CUS_Q_13	7586	3.70	0.98	0.55

ASD scale

Item	N responses	Mean response	SD	Item-total cor.
V2_ASD_Q_1	7609	3.74	0.97	0.37
V2_ASD_Q_2	7602	2.83	1.28	0.39
V2_ASD_Q_3	7618	2.94	1.22	0.38

AAL scale

This scale included one field test item, item 17. First, we provide statistics including this item:

Item	N responses	Mean response	SD	Item-total cor.
V2_AAL_Q_1	7607	2.69	1.15	0.52
V2_AAL_Q_2	7598	3.13	1.15	0.53
V2_AAL_Q_15	7600	3.44	1.06	0.52
V2_AAL_Q_17	843	3.40	0.93	0.56

Finally, statistics with the field test item excluded:

Item	N responses	Mean response	SD	Item-total cor.
V2_AAL_Q_1	7607	2.69	1.15	0.44
V2_AAL_Q_2	7598	3.13	1.15	0.48
V2_AAL_Q_15	7600	3.44	1.06	0.39

Appendix E: EFA of all Pulse Check Scale items combined with one to eleven factors

Total variance explained and model fit measures for all solutions

# Factors	Prop. variance expl.	RMSEA	RMSR	TLI
1	0.38	0.07	0.04	0.86
2	0.41	0.06	0.04	0.89
3	0.43	0.05	0.03	0.91
4	0.45	0.05	0.03	0.92
5	0.46	0.05	0.02	0.93
6	0.48	0.04	0.02	0.94
7	0.49	0.04	0.02	0.95
8	0.50	0.04	0.01	0.96
9	0.51	0.03	0.01	0.97
10	0.52	0.03	0.01	0.98
11	0.54	0.02	0.01	0.99

Note: RMSEA: root mean square error of approximation. RMSR: root mean square residual. TLI: Tucker-Lewis index.

Loadings: one factor

	PA1
V2_HE_Q_1	0.6
V2_HE_Q_2	0.63
V2_HE_Q_3	0.7
V2_HE_Q_13	0.65
V2_WCF_Q_1	0.69
V2_WCF_Q_2	0.71
V2_RL_Q_1	0.67
V2_RL_Q_2	0.63
V2_RL_Q_7	0.71
V2_REL_Q_1	0.58
V2_REL_Q_2	0.66
V2_REL_Q_3	0.63
V2_ASO_Q_1	0.58
V2_ASO_Q_2	0.65
V2_ASO_Q_3	0.62
V2_SCA_Q_16	0.41
V2_SCA_Q_17	0.57
V2_CC_Q_1	0.68
V2_CC_Q_2	0.66
V2_CUS_Q_1	0.57
V2_CUS_Q_2	0.67
V2_CUS_Q_8	0.62
V2_CUS_Q_13	0.63
V2_ASD_Q_1	0.64
V2_ASD_Q_2	0.49
V2_ASD_Q_3	0.51
V2_AAL_Q_1	0.54
V2_AAL_Q_2	0.55
V2_AAL_Q_15	0.48

Loadings: two factors

	PA1	PA2
V2_HE_Q_1	0.23	0.41
V2_HE_Q_2	0.4	0.26
V2_HE_Q_3	0.46	0.27
V2_HE_Q_13	0.34	0.35
V2_WCF_Q_1	0.32	0.41
V2_WCF_Q_2	0.35	0.41
V2_RL_Q_1	0.48	0.22
V2_RL_Q_2	0.64	
V2_RL_Q_7	0.57	
V2_REL_Q_1	0.68	
V2_REL_Q_2	0.68	
V2_REL_Q_3	0.47	
V2_ASO_Q_1	0.47	
V2_ASO_Q_2		0.94
V2_ASO_Q_3		0.77
V2_SCA_Q_16	0.48	
V2_SCA_Q_17	0.5	
V2_CC_Q_1		0.73
V2_CC_Q_2		0.79
V2_CUS_Q_1	0.58	
V2_CUS_Q_2	0.4	0.31
V2_CUS_Q_8	0.39	0.26
V2_CUS_Q_13	0.42	0.24
V2_ASD_Q_1	0.31	0.37
V2_ASD_Q_2	0.41	
V2_ASD_Q_3	0.64	
V2_AAL_Q_1	0.67	
V2_AAL_Q_2	0.49	
V2_AAL_Q_15	0.39	

Loadings: three factors

	PA1	PA3	PA2
V2_HE_Q_1		0.49	0.23
V2_HE_Q_2	0.28	0.21	
V2_HE_Q_3	0.31	0.28	
V2_HE_Q_13		0.31	0.23
V2_WCF_Q_1	0.42		0.43
V2_WCF_Q_2	0.48		0.45
V2_RL_Q_1		0.48	
V2_RL_Q_2	0.5	0.2	
V2_RL_Q_7	0.36	0.34	
V2_REL_Q_1	0.55		
V2_REL_Q_2	0.5	0.26	
V2_REL_Q_3		0.49	
V2_ASO_Q_1	0.49		
V2_ASO_Q_2			0.84
V2_ASO_Q_3			0.73
V2_SCA_Q_16	0.43		
V2_SCA_Q_17	0.44		
V2_CC_Q_1			0.64
V2_CC_Q_2			0.69
V2_CUS_Q_1	0.33	0.38	
V2_CUS_Q_2		0.72	
V2_CUS_Q_8		0.78	
V2_CUS_Q_13		0.68	
V2_ASD_Q_1		0.42	0.22
V2_ASD_Q_2	0.43		
V2_ASD_Q_3	0.54		
V2_AAL_Q_1	0.81		
V2_AAL_Q_2	0.56		
V2_AAL_Q_15	0.31		

Loadings: four factors

	PA3	PA1	PA2	PA4
V2_HE_Q_1	0.49		0.22	
V2_HE_Q_2	0.21			
V2_HE_Q_3	0.27	0.23		
V2_HE_Q_13	0.31		0.24	
V2_WCF_Q_1		0.24	0.44	
V2_WCF_Q_2		0.31	0.46	
V2_RL_Q_1	0.47			
V2_RL_Q_2		0.52		
V2_RL_Q_7	0.33	0.25		
V2_REL_Q_1		0.51		
V2_REL_Q_2	0.25	0.47		
V2_REL_Q_3	0.49			
V2_ASO_Q_1		0.46		
V2_ASO_Q_2			0.87	
V2_ASO_Q_3			0.77	
V2_SCA_Q_16				0.66
V2_SCA_Q_17				0.78
V2_CC_Q_1			0.66	
V2_CC_Q_2			0.71	
V2_CUS_Q_1	0.38	0.46		
V2_CUS_Q_2	0.71			
V2_CUS_Q_8	0.77			
V2_CUS_Q_13	0.67			
V2_ASD_Q_1	0.41		0.22	
V2_ASD_Q_2		0.49		
V2_ASD_Q_3		0.6		
V2_AAL_Q_1		0.72		
V2_AAL_Q_2		0.58		
V2_AAL_Q_15		0.37		

Loadings: five factors

	PA3	PA2	PA1	PA4	PA5
V2_HE_Q_1	0.52				
V2_HE_Q_2	0.29		0.28		
V2_HE_Q_3	0.26				
V2_HE_Q_13	0.33	0.2			
V2_WCF_Q_1		0.4	0.3		
V2_WCF_Q_2		0.42	0.35		
V2_RL_Q_1	0.51				
V2_RL_Q_2			0.26		0.31
V2_RL_Q_7	0.33				
V2_REL_Q_1	0.2		0.48		
V2_REL_Q_2	0.29		0.43		
V2_REL_Q_3	0.63		0.36		
V2_ASO_Q_1			0.34		
V2_ASO_Q_2		0.85			
V2_ASO_Q_3		0.81			
V2_SCA_Q_16				0.72	
V2_SCA_Q_17				0.75	
V2_CC_Q_1		0.64			
V2_CC_Q_2		0.68			
V2_CUS_Q_1	0.24				0.56
V2_CUS_Q_2	0.68				
V2_CUS_Q_8	0.72				
V2_CUS_Q_13	0.61				0.21
V2_ASD_Q_1	0.46				
V2_ASD_Q_2			0.27		0.25
V2_ASD_Q_3					0.51
V2_AAL_Q_1			0.69		
V2_AAL_Q_2			0.71		
V2_AAL_Q_15			0.47		

Loadings: six factors

	PA2	PA1	PA3	PA4	PA6	PA5
V2_HE_Q_1		0.5				
V2_HE_Q_2						
V2_HE_Q_3		0.26				
V2_HE_Q_13	0.24	0.25				
V2_WCF_Q_1	0.45					
V2_WCF_Q_2	0.46		0.2			
V2_RL_Q_1		0.45				
V2_RL_Q_2						0.36
V2_RL_Q_7		0.3				
V2_REL_Q_1					0.42	
V2_REL_Q_2					0.84	
V2_REL_Q_3		0.41			0.58	
V2_ASO_Q_1	0.21					0.2
V2_ASO_Q_2	0.89					
V2_ASO_Q_3	0.84					
V2_SCA_Q_16				0.79		
V2_SCA_Q_17				0.84		
V2_CC_Q_1	0.71					
V2_CC_Q_2	0.81					
V2_CUS_Q_1		0.24				0.62
V2_CUS_Q_2		0.59				
V2_CUS_Q_8		0.64				
V2_CUS_Q_13		0.6				0.23
V2_ASD_Q_1		0.42	0.2			
V2_ASD_Q_2						0.3
V2_ASD_Q_3						0.57
V2_AAL_Q_1			0.5			
V2_AAL_Q_2			0.74			
V2_AAL_Q_15		0.22	0.73			

Loadings: seven factors

	PA2	PA1	PA5	PA6	PA7	PA4	PA3
V2_HE_Q_1		0.43					
V2_HE_Q_2							
V2_HE_Q_3		0.24					
V2_HE_Q_13	0.25	0.24					
V2_WCF_Q_1	0.42						
V2_WCF_Q_2	0.43						0.2
V2_RL_Q_1		0.36					0.28
V2_RL_Q_2			0.49				
V2_RL_Q_7		0.22	0.24				0.27
V2_REL_Q_1				0.41			
V2_REL_Q_2				0.84			
V2_REL_Q_3		0.3		0.67			0.25
V2_ASO_Q_1			0.23				
V2_ASO_Q_2	0.84						
V2_ASO_Q_3	0.79						
V2_SCA_Q_16						0.72	
V2_SCA_Q_17						0.82	
V2_CC_Q_1	0.74						
V2_CC_Q_2	0.85						
V2_CUS_Q_1		0.3	0.61				
V2_CUS_Q_2		0.55					
V2_CUS_Q_8		0.62					
V2_CUS_Q_13		0.58					
V2_ASD_Q_1		0.33					0.33
V2_ASD_Q_2			0.38				
V2_ASD_Q_3			0.64				
V2_AAL_Q_1					0.57		
V2_AAL_Q_2					0.73		
V2_AAL_Q_15					0.66		

Loadings: eight factors

	PA1	PA2	PA7	PA4	PA6	PA5	PA3	PA8
V2_HE_Q_1	0.53							
V2_HE_Q_2	0.25		0.32					
V2_HE_Q_3	0.32							
V2_HE_Q_13	0.35		0.21					
V2_WCF_Q_1			0.79					
V2_WCF_Q_2			0.73					
V2_RL_Q_1	0.45		0.25					
V2_RL_Q_2			0.31			0.43		
V2_RL_Q_7	0.34		0.5					
V2_REL_Q_1					0.4			
V2_REL_Q_2					0.72			
V2_REL_Q_3	0.27				0.76			0.25
V2_ASO_Q_1						0.21		
V2_ASO_Q_2		0.81						0.21
V2_ASO_Q_3		0.71						
V2_SCA_Q_16							0.96	
V2_SCA_Q_17			0.29				0.42	
V2_CC_Q_1		0.55						
V2_CC_Q_2		0.64						
V2_CUS_Q_1	0.36					0.55		
V2_CUS_Q_2	0.65							
V2_CUS_Q_8	0.73							
V2_CUS_Q_13	0.71							
V2_ASD_Q_1	0.35							0.29
V2_ASD_Q_2						0.36		
V2_ASD_Q_3						0.59		
V2_AAL_Q_1				0.54				
V2_AAL_Q_2				0.73				
V2_AAL_Q_15	0.2			0.64				

Loadings: nine factors

	PA1	PA9	PA8	PA6	PA5	PA7	PA4	PA2	PA3
V2_HE_Q_1	0.56								
V2_HE_Q_2	0.26								
V2_HE_Q_3	0.32								
V2_HE_Q_13	0.34					0.37			
V2_WCF_Q_1			0.84						
V2_WCF_Q_2			0.76						
V2_RL_Q_1	0.5								
V2_RL_Q_2		0.56							
V2_RL_Q_7	0.39	0.25	0.31						
V2_REL_Q_1					0.37				
V2_REL_Q_2					0.73				
V2_REL_Q_3	0.33				0.68				
V2_ASO_Q_1		0.33							
V2_ASO_Q_2								0.69	
V2_ASO_Q_3								0.61	
V2_SCA_Q_16							0.8		
V2_SCA_Q_17							0.62		
V2_CC_Q_1						0.58		0.22	
V2_CC_Q_2						0.67		0.27	
V2_CUS_Q_1	0.38	0.35							0.45
V2_CUS_Q_2	0.69								
V2_CUS_Q_8	0.76								
V2_CUS_Q_13	0.78								0.24
V2_ASD_Q_1	0.4								
V2_ASD_Q_2		0.66							
V2_ASD_Q_3		0.68							
V2_AAL_Q_1				0.55					
V2_AAL_Q_2				0.71					
V2_AAL_Q_15	0.25			0.6					

Loadings: ten factors

	PA1	PA9	PA10	PA7	PA6	PA4	PA2	PA5	PA8	PA3
V2_HE_Q_1	0.51									
V2_HE_Q_2			0.2					0.25		
V2_HE_Q_3	0.26									
V2_HE_Q_13	0.27		0.41							
V2_WCF_Q_1				0.84						
V2_WCF_Q_2				0.72						
V2_RL_Q_1	0.36							0.42		
V2_RL_Q_2								0.5	0.3	
V2_RL_Q_7								0.7		
V2_REL_Q_1					0.37					
V2_REL_Q_2					0.73					
V2_REL_Q_3	0.37				0.67					
V2_ASO_Q_1									0.26	
V2_ASO_Q_2							0.71			
V2_ASO_Q_3							0.68			
V2_SCA_Q_16						0.85				
V2_SCA_Q_17						0.6				
V2_CC_Q_1			0.62				0.2			
V2_CC_Q_2			0.71				0.27			
V2_CUS_Q_1	0.2									0.53
V2_CUS_Q_2	0.57									0.21
V2_CUS_Q_8	0.65									0.23
V2_CUS_Q_13	0.64									0.31
V2_ASD_Q_1	0.46								0.23	
V2_ASD_Q_2									0.6	
V2_ASD_Q_3									0.43	0.25
V2_AAL_Q_1		0.62								
V2_AAL_Q_2		0.72								
V2_AAL_Q_15	0.25	0.62								

Loadings: eleven factors

	PA1	PA10	PA11	PA3	PA8	PA6	PA4	PA5	PA9	PA2	PA7
V2_HE_Q_1	0.31	0.21									0.24
V2_HE_Q_2								0.98			
V2_HE_Q_3	0.21										
V2_HE_Q_13	0.2	0.3						0.26			
V2_WCF_Q_1				0.83							
V2_WCF_Q_2				0.78							
V2_RL_Q_1	0.22								0.34		
V2_RL_Q_2			0.39						0.46		
V2_RL_Q_7									0.83		
V2_REL_Q_1						0.39					
V2_REL_Q_2						0.76					
V2_REL_Q_3						0.68					0.28
V2_ASO_Q_1			0.32								
V2_ASO_Q_2										0.67	
V2_ASO_Q_3										0.67	
V2_SCA_Q_16							0.9				
V2_SCA_Q_17							0.53				
V2_CC_Q_1		0.76									
V2_CC_Q_2		0.72								0.23	
V2_CUS_Q_1	0.52		0.36								
V2_CUS_Q_2	0.6										
V2_CUS_Q_8	0.7										
V2_CUS_Q_13	0.76										
V2_ASD_Q_1			0.22								0.44
V2_ASD_Q_2			0.67								0.22
V2_ASD_Q_3			0.65								
V2_AAL_Q_1					0.55						
V2_AAL_Q_2					0.71						
V2_AAL_Q_15					0.63						

Appendix F: Results of ANOVAs with School and Interacted Demographic Predictors for All Pulse Check Leaps

HE scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	28	402.76	14.38	25.20	0.00	10.49
Race	5	3.68	0.74	1.29	0.27	0.10
Gender	1	0.15	0.15	0.27	0.60	0.00
FRL	2	0.10	0.05	0.09	0.92	0.00
Race x Gender	5	1.99	0.40	0.70	0.63	0.05
Race x FRL	3	1.05	0.35	0.61	0.61	0.03
Gender x FRL	1	0.10	0.10	0.17	0.68	0.00
Race x Gender x FRL	3	0.18	0.06	0.10	0.96	0.00
Student/Residual	6006	3428.78	0.57	–	–	89.32

WCF scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	4	64.20	16.05	19.94	0.00	5.70
Race	3	0.83	0.28	0.35	0.79	0.07
Gender	1	0.68	0.68	0.84	0.36	0.06
Race x Gender	3	0.57	0.19	0.24	0.87	0.05
Student/Residual	1316	1059.22	0.80	–	–	94.11

Note: FRL was excluded due to lack of variance in FRL among students who took all WCF items.

RL scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	28	299.68	10.70	16.66	0.00	7.11
Race	5	1.12	0.22	0.35	0.88	0.03
Gender	1	0.01	0.01	0.01	0.92	0.00
FRL	2	2.14	1.07	1.67	0.19	0.05
Race x Gender	5	4.12	0.82	1.28	0.27	0.10
Race x FRL	3	0.31	0.10	0.16	0.92	0.01

Gender x FRL	1	0.00	0.00	0.00	1.00	0.00
Race x Gender x FRL	3	1.68	0.56	0.87	0.45	0.04
Student/Residual	6081	3907.31	0.64	–	–	92.67

REL scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	28	482.04	17.22	20.68	0.00	8.64
Race	5	3.47	0.69	0.83	0.53	0.06
Gender	1	5.96	5.96	7.16	0.01	0.11
FRL	2	1.13	0.56	0.68	0.51	0.02
Race x Gender	5	12.05	2.41	2.90	0.01	0.22
Race x FRL	3	0.22	0.07	0.09	0.97	0.00
Gender x FRL	1	2.37	2.37	2.85	0.09	0.04
Race x Gender x FRL	3	0.54	0.18	0.22	0.89	0.01
Student/Residual	6090	5069.07	0.83	–	–	90.89

ASO scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	2	12.71	6.35	8.14	0.00	2.15
Race	4	2.82	0.70	0.90	0.46	0.48
Gender	1	16.73	16.73	21.44	0.00	2.83
FRL	1	0.35	0.35	0.45	0.50	0.06
Race x Gender	3	0.62	0.21	0.27	0.85	0.11
Race x FRL	3	3.25	1.08	1.39	0.24	0.55
Gender x FRL	1	0.05	0.05	0.06	0.80	0.01
Race x Gender x FRL	2	0.44	0.22	0.28	0.76	0.07
Student/Residual	711	554.89	0.78	–	–	93.75

SCA scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance

School	6	108.57	18.10	21.03	0.00	6.96
Race	5	2.88	0.58	0.67	0.65	0.18
Gender	1	3.87	3.87	4.50	0.03	0.25
FRL	1	0.56	0.56	0.65	0.42	0.04
Race x Gender	4	4.05	1.01	1.18	0.32	0.26
Race x FRL	1	0.03	0.03	0.03	0.85	0.00
Gender x FRL	1	0.00	0.00	0.00	0.98	0.00
Race x Gender x FRL	1	0.21	0.21	0.24	0.62	0.01
Student/Residual	1673	1439.46	0.86	–	–	92.29

CC scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	28	568.10	20.29	21.12	0.00	8.75
Race	5	4.41	0.88	0.92	0.47	0.07
Gender	1	10.33	10.33	10.75	0.00	0.16
FRL	2	0.38	0.19	0.20	0.82	0.01
Race x Gender	5	2.91	0.58	0.61	0.70	0.04
Race x FRL	3	1.17	0.39	0.41	0.75	0.02
Gender x FRL	1	2.30	2.30	2.40	0.12	0.04
Race x Gender x FRL	3	5.42	1.81	1.88	0.13	0.08
Student/Residual	6135	5894.33	0.96	–	–	90.83

CUS scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	28	261.18	9.33	17.00	0.00	7.33
Race	5	3.17	0.63	1.15	0.33	0.09
Gender	1	0.21	0.21	0.38	0.54	0.01
FRL	2	0.41	0.20	0.37	0.69	0.01
Race x Gender	5	4.00	0.80	1.46	0.20	0.11
Race x FRL	3	0.07	0.02	0.04	0.99	0.00
Gender x FRL	1	0.61	0.61	1.12	0.29	0.02

Race x Gender x FRL	3	0.40	0.13	0.24	0.87	0.01
Student/Residual	6002	3293.92	0.55	–	–	92.42

ASD scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	28	184.65	6.59	9.25	0.00	4.06
Race	5	1.18	0.24	0.33	0.89	0.03
Gender	1	0.03	0.03	0.05	0.83	0.00
FRL	2	0.16	0.08	0.11	0.89	0.00
Race x Gender	5	5.13	1.03	1.44	0.21	0.11
Race x FRL	3	0.86	0.29	0.40	0.75	0.02
Gender x FRL	1	1.55	1.55	2.17	0.14	0.03
Race x Gender x FRL	3	2.46	0.82	1.15	0.33	0.05
Student/Residual	6101	4350.19	0.71	–	–	95.69

AAL scale

Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	% variance
School	3	22.34	7.45	13.44	0.00	5.32
Race	1	2.82	2.82	5.08	0.02	0.67
Gender	4	2.66	0.66	1.20	0.31	0.63
Race x Gender	4	1.54	0.38	0.69	0.60	0.37
Student/Residual	704	390.16	0.55	–	–	93.00

Note: FRL was excluded due to lack of variance in FRL among students who took all AAL items.