Motivation, Engagement, and Beliefs Survey Validation Report

September 2016

Neil Naftzger

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

7525_09/16

Contents

Introduction

In recent years, an increasing level of attention has been paid to the importance of supporting youth in developing attitudes, beliefs, and skills critical to success in school and life that go beyond the academic content students learn as they proceed through the PK–12 educational system. This set of positive youth outcomes are typically described under the mantle of social and emotional learning (SEL), character development, and noncognitive outcomes and include things like self-regulation, positive mindsets, persistence, and interpersonal skills. Within the workforce development realm, these outcomes are commonly referred to as soft skills.

During the past several years, an increasing degree of attention has been paid to the role youth development and other out-of-school-time and afterschool programs can have in supporting the development of these types of attitudes, beliefs, and skills (Devaney, 2015a; Durlak, Weissberg, & Pachan, 2010; Moroney, 2016; Smith, McGovern, Larson, Hillaker, & Peck, 2016). An important feature of these programs is that they are particularly well designed to provide youth, particularly as they enter adolescence, with enhanced opportunities to experience agency—to use their growing cognitive, reasoning, and regulatory capacities to interact with increasingly complex systems to achieve desired goals (Larson & Angus, 2011). Programmatic approaches like project- and inquiry-based learning commonly found in out-of-school settings, offer a fertile platform from which to support the cultivation of key attitudes, beliefs, and skills needed for success in the 21st century.

In addition, the OST field in particular has been very proactive in constructing and adopting quality assessment tools and improvement systems oriented at helping activity leaders learn and adopt practices that ensure youth experience a supportive, interactive, and engaging learning environment (the Youth Program Quality Assessment and the Assessment of Program Practices Tool are examples) and participate in skill-building-oriented activities in areas linked to the promotion of social and emotional learning (Smith et al., 2016). Opportunities to experience relatedness, a sense of belonging, and collective flow have been identified as critical to not only supporting youth motivation, interest, and engagement in out-of-school activities (Larson & Dawes, in press) but also serve to support the development of youth SEL-related competencies (Devaney, 2015b; Smith et al., 2016).

Although this type of outcome is increasingly gaining traction in the educational and workforce development fields as key determinants of youth success (Farrington et al., 2012; Wilson-Ahlstrom, Yohalem, DuBois, Ji, & Hillaker, 2014), efforts to measure youth development in these areas within the confines of afterschool programs are still in their infancy. This is likely the case for three primary reasons:

1. There is uncertainty in the field about what exactly to measure and how best to measure it.
2. There is a sense that efforts to assess youth development on these outcomes are somehow inferior to assessing program impact on academic outcomes, such as test results in reading and mathematics.
3. There are concerns that results from these measurement practices objectify youth and can be used to stigmatize certain populations and the communities they come from.

Although these three reasons are still valid and present, efforts are increasingly underway to develop better measures, clarify the relationship between these types of outcomes and academic achievement, and determine the best use of these data to describe program outcomes and inform quality improvement efforts without stigmatizing certain populations. One such example is the *Motivation, Engagement, and Beliefs Survey* developed by the Youth Development Executives of King County (YDEKC).

Since 2012, YDEKC has been working with youth development agencies in King County, Washington, and elsewhere to both articulate the manner in which these agencies are likely having an impact on youth in these critical areas and work to develop measures that afford these agencies a better sense of how youth are functioning in these areas and potentially developing in ways that are likely to support their success in school and life more broadly. American Institutes for Research (AIR) began working with YDEKC in 2013 in order to assess the functioning of the *Motivation, Engagement, and Beliefs Survey* and to modify the tool for use in its efforts to evaluate the Washington 21st Century Community Learning Centers (21st CCLC) program.

The purpose of this report is summarize the evidence AIR has collected and analyzed pertaining to the reliability and validity of the version of *Motivation, Engagement, and Beliefs Survey* used as part of the evaluation of the Washington 21st CCLC program. First, steps will be taken to briefly summarize AIR's involvement in modifying and using the tool and outline the scales appearing on the version that was used by AIR during the 2014–15 and 2015–16 school years in supporting evaluation activities related to the Washington 21st CCLC program. Next, information about the sample and methods used to review the functioning of the tool are detailed, results are summarized, and conclusions about the reliability and validity of the *Motivation, Engagement, and Beliefs Survey* are provided.

Background

Starting in fall 2013, AIR began working with YDEKC and 21st CCLC program staff at the Washington Office of the Superintendent of Public Instruction (OSPI) to further refine the tool for use with the state's 21st CCLC programs. The goal of these efforts was to improve the psychometric functioning of the tool, add items that provided youth with the opportunity to reflect on how they had benefited from program participation, and provide centers funded by 21st CCLC with information about how the youth they served were doing on key attitudes, beliefs, and skills related to SEL and noncognitive factors linked to school and work success. A revised version of the tool with new items pertaining to perceived program impact was piloted in spring 2014 in 38 centers funded by 21st CCLC, resulting in 1,199 completed surveys taken by youth in Grades 4 to 12. Results from the 2014 pilot resulted in a series of substantial modifications to the survey, including the following:

- A reduction in the number of constructs being measured, going from 14 to seven scales represented on the postpilot version of the tool.
- A substantial reassignment of items to scales represented in the new structure
- A revision in the rating scale to move from a five-point to a four-point scale
- The elimination of 23 items from the survey and the addition of two new items
- Changes in phrasing for some items

The revised version of the *Motivation, Engagement, and Beliefs Survey* was administered in spring 2015 in all 21st CCLC programs serving youth in Grades 4–12. A total of 4,968 completed surveys were collected during spring 2015 from 141 21st CCLC programs. An assessment of how well the revised survey performed led to the elimination of the three additional items from the survey. The final structure of the survey then is composed of 44 items on seven scales.

Survey Overview

Three types of scales are found on the final version of the *Youth Motivation, Engagement, and Beliefs Survey*. A full copy of the survey can be found in Appendix A.

1. *Items pertaining to youth's sense of belonging and engagement in the 21st CCLC program*. The purpose of these items was to obtain authentic feedback from youth on their experiences in the 21st CCLC program they were enrolled in during the school year. Examples of items of this type included *I feel proud to be part of my program*; *This program helps me build new skills*; and *What we do in this program is challenging in a good way*. For all items appearing on the survey, youth were asked to respond to each item by endorsing one of the following response options*:* "not at all true"; *"*somewhat true*"*; "mostly true"; or "completely true".
2. *Items pertaining to youth's sense of how they may have been affected by participation in the program*. The purpose of these items was to explore the extent to which youth believed the program may have helped them in terms of developing positive academic behaviors and better self-management skills. Examples of items of this type included *This program has helped me to become more interested in what I'm learning in school* and *This program has helped me get better at staying focused on my work.*
3. *Items pertaining to how youth reported functioning at present when taking the survey on a series of areas related to positive youth development*. The purpose of these items was to gauge how well youth described themselves as doing in four key areas: (a) academic identity, (b) positive mindsets, (c) self-management, and (d) interpersonal skills. Examples of items appearing on these scales include *Doing well in school is an important part of who I am* (academic identity); *I can solve difficult problems if I try hard enough* (mindsets); *I can calm myself down when I'm excited or upset* (self-management); and *I work well with others on group projects* (interpersonal skills).

In this sense, the *Youth Motivation, Engagement, and Beliefs Survey* is meant to collect data to serve four primary purposes:

1. Assess how well the program has established a good fit with youth served in the program and how successful they have been in constructing a setting where youth see value in participating in program activities
2. Gain information from youth on how they may have benefited from participation in program activities

3. Assess how well youth may be functioning on a series of constructs related to positive social and emotional development and mindsets and attitudes related to success in school and in life more broadly
4. Assess the extent to which youth grow on the aforementioned constructs during a programming year

A primary goal of this report is to explore the extent to which each of these purposes is served on the basis of how well the tool was found to function across a series of analyses oriented to assessing the reliability and validity of the tool.

Approach and Sample

In order to explore the underlying reliability and validity of the scales appearing on the *Youth Motivation, Engagement, and Beliefs Survey,* a series of analyses were undertaken relying on survey data collected in spring 2015 and spring 2016 as part of the statewide evaluation of the Washington 21st CCL program. During both data collection periods, all centers funded by 21st CCLC serving youth enrolled in Grades 4–12 were required to submit youth survey data for those youth enrolled in these grade levels who attended programming 30 days or more during the school year and were still accessible to take the survey.

A total of 4,497 were collected from 139 centers in spring 2015, and 3,750 from 121 centers in spring 2016 in the targeted grade levels. As shown in Table 1, more than 80 percent of completed surveys were taken by youth in Grades 4–8, with the majority of respondents in Grades 4–6. In each year, 8 to 10 percent of completed survey were missing grade-level information on the respondent. Surveys with missing grade-level information were retained for the analyses summarized in this report because date-of-birth information was provided for these respondents. Youth who were nine years old or older at the start of the school year in question were retained as part of the study sample.

Table 1. Summary of Survey Respondents by Grade Level and Year

	Spring 2015		Spring 2016	
	N	**%**	***N***	**%**
4th grade	929	20.7%	920	24.5%
5th grade	795	17.7%	817	21.8%
6th grade	899	20.0%	666	17.8%
7th grade	633	14.1%	423	11.3%
8th grade	493	11.0%	358	9.5%
9th grade	189	4.2%	44	1.2%
10th grade	81	1.8%	65	1.7%
11th grade	68	1.5%	55	1.5%
12th grade	58	1.3%	38	1.0%
Missing	352	7.8%	364	9.7%
Total	4,497	100%	3,750	100%

As shown in Table 2, the average number of days youth responding to the survey attended 21st CCLC programming was slightly more than 70 days during the 2014–15 and 2015–16 school years.

Table 2. Mean Number of Days of 21st CCLC Programming Attended During the School Year

	Spring 2015	Spring 2016
Mean SY Days Attended	70.8	71.6

Approach

In order to assess how well the revised version of the *Youth Motivation, Engagement, and Beliefs Survey* was functioning in measuring the domain of constructs under consideration, a series of analyses was undertaken to examine various facets of reliability and validity oriented to answering the following set of questions.

1. ***Are survey scales functioning well psychometrically?*** This research question was addressed by conducting a series of analyses employing the Rasch rating scale model in Winsteps to explore the following validity facets: (1) content validity (technical quality of the items); (2) structural validity (unidimensionality); (3) substantive validity (rating scale functioning); (4) generalizability (reliability); and (5) external validity (responsiveness/sensitivity; group comparisons). This approach was informed by the instrument validation approach described by Wolfe & Smith (2007). These analyses were conducted using both the 2015 and 2016 samples in order to assess if results were consistent across the two samples.
2. ***Is the level of youth functioning on survey scales predictive of school-related outcomes in the manner hypothesized?*** This question addresses the question of predictive validity (a facet of criterion-related validity)—are scores on survey constructs related to youth performance on school-related outcomes during the school year in question? Answering this question is important for understanding whether survey subscales can serve as viable proxies for youth performance on key school-related outcomes, including academic achievement and behaviors like school attendance and disciplinary incidents. School-related outcome data was obtained directly from OSPI as part of the statewide evaluation of the 21st CCLC program. The data available for these analyses were from the 2015 sample for school-related outcomes associated with the 2014–15 school year.

A summary of each facet of validity examined and the approaches undertaken to assess each validity component are summarized in greater detail in Table 3.

Table 3. Summary of Validity Aspects to Be Examined and Summary of Approach

Validity Aspect	Validity Subelement	Questions Addressed	Rasch Statistics/ Approaches
Content validity	Technical quality of the items	➢ To what extent is there a discrepancy between the expected and observed values associated with a particular item or subset of items? ➢ To what extent are scores on a particular item consistent with the average score across the remaining items making up the subscale in question?	➢ Mean-squared outfit indices within acceptable range ➢ Point measure correlations within acceptable range.
Structural validity	Unidimensionality	➢ Do the items making up a given subscale appear to be underpinned by a single latent construct?	➢ Principal component analysis of the standardized residuals resulting from the Rasch scaling of survey scale data indicates the absence of a second major factor.
Substantive validity	Rating scale functioning	➢ To what extent does the rating scale underpinning the survey scale behave in a manner that is consistent with Rasch guidelines related to how rating scales should function?	➢ Each rating scale category contains a minimum of 10 responses. ➢ Smooth and unimodal shape to each rating scale distribution; structure calibrations increase monotonically. ➢ Average respondent measure for each category increases monotonically. ➢ Unweighted mean-squared fit indices within acceptable range.
Generalizability	Reliability	➢ How consistent are survey scale-derived measures across a variety of measurement facets?	➢ Cronbach's alpha and person separation reliability indices are within an acceptable range.

Validity Aspect	Validity Subelement	Questions Addressed	Rasch Statistics/ Approaches
External validity	Responsiveness/ sensitivity	➢ Do the distribution of the item calibrations and rating scale thresholds show that survey scale-derived measures likely to be sensitive to detecting youth change?	➢ Distribution of item calibrations and rating scale thresholds assessed via offering-item maps suggest program change is detectable. Absence of ceiling or floor effects. Person strata index indicates it is possible to distinguish between multiple levels of a trait.
	Group and within-individual comparisons	➢ To what extent do youth, especially youth scoring lower on survey scales on the presurvey, demonstrate growth on survey measures across time? To what extent is the level of growth associated with youth experiences in programming?	➢ Matched sample *t*-tests and measure correlations
Criterion-related Validity	Predictive validity	➢ Is youth functioning on survey subscales related to school-related outcomes in the manner predicted?	➢ Correlative analyses employing hierarchical linear modeling

Are Survey Scales Functioning Well Psychometrically? Content and Structural Validity

Item Fit

When conducting analyses employing the Rasch rating scale model, the concept of mean-squared outfit is used to assess how well an item fits within a particular measure subscale (that is, should an item be included in the scale?). Using information about how an individual youth scored across the full domain of items on the scale in question and how the full sample of youth scored on a given item, the Rasch model constructed an expected score for each youth on each item represented in the analysis. This expected score was then compared with the observed score for the youth on the item in question and a residual was calculated. The value of the mean squared residual among the items represented in an analysis serves as an indication of how well

the item in question fit within the subscale under consideration. According to Linacre (2009), outfit mean-square fit statistics should be between 0.5 and 1.5 for an item if it can be deemed usable for the purposes of measurement. As show in Table 4, this was found to be the case for the full domain of items represents in each of the scales appearing on the *Youth Motivation, Engagement, and Beliefs Survey* in both the 2015 and the 2016 sample.

In addition to the mean outfit values, the technical quality of the items making up a subscale also can be assessed by examining the extent to which scores on an item are consistent with the average score across the remaining items making up the subscale in question. In Winsteps, this relationship is assesses by the point measure correlation. According to Wolfe and Smith (2007), ideally an item associated with a polytomous rating scale with have a point measure correlation value that is greater than .40. As shown in Table 4, this was found to be the case with all items across all subscales across both sample years.

From the results shown in Table 4, we can generally conclude that the items making up each scale appearing on the *Youth Motivation, Engagement, and Beliefs Survey* fit well within the scales they are associated with, providing positive evidence of the content validity of survey scales from a psychometric perspective.

Table 4. Range of Item-Level Outfit Mean-Square Fit Statistics and Point Measure Correlations by Subscale

	Outfit Mean-Square Fit Statistics		Point Measure Correlation	
Survey Subscale	**2015**	**2016**	**2015**	**2016**
Academic identity	.92 to 1.06	.90 to 1.07	.74 to .79	.77 to .81
Positive mindsets	.85 to 1.11	.80 to 1.12	.62 to .68	.63 to .70
Self-management	.85 to 1.17	.87 to 1.20	.59 to .68	.64 to .67
Interpersonal skills	.91 to 1.17	.87 to 1.23	.63 to .70	.64 to .72
Program impact—academics	.81 to 1.20	.82 to 1.17	.78 to .84	.80 to .85
Program impact—self-management	.92 to 1.08	.91 to 1.07	.74 to .80	.77 to .82
Program belonging and engagement	.90 to 1.10	.89 to 1.10	.76 to .80	.77 to .81

Unidimensionality

The Rasch model is predicated on an assumption of unidimensionality in the measure, meaning all items on a subscale should be associated with a single latent construct. In Winsteps, the unidimensionality of a survey subscale was assessed by performing a principal component analysis of the residuals resulting from the calibration of the subscale in question. As noted by Linacre (2009), the principal component analysis of the residuals performed in Winsteps demonstrates the contrast between opposing factors as opposed to loadings on a single factor. In this regard, the principal component analysis procedure yields information about the amount of variance explained by the dimension formulated by application of the Rasch rating scale model, the amount of that variation explained by the items making up the scale in question, and finally, the amount of

variation accounted by a potential second dimension. In this regard, the Rasch model fits the data to a unidimensional model and then looks for structure in the residuals variance.

Generally, unidimensionality can be considered to be retained if the amount of variation explained by the biggest secondary dimension has the strength of less than two items (that is, an eigenvalue less than 2.00). As shown in Table 5, this was consistently found to be the case in relation to each subscale across both samples. Overall, these results are supportive of the unidimensional nature of each scale.

Table 5. Percentage of Variation Explained by Rasch Measures and Eigenvalues Associated With the Largest Second Factor by Subscale

	% of Variation Explained by the Measure		Eigenvalue Associated With the Largest Second Factor	
Survey Subscale	**2015**	**2016**	**2015**	**2016**
Academic identity	50.7%	52.0%	1.4	1.4
Positive mindsets	44.1%	45.3%	1.5	1.5
Self-management	39.8%	41.4%	1.6	1.6
Interpersonal skills	43.0%	46.3%	1.4	1.4
Program impact—academics	55.3%	57.3%	1.4	1.4
Program impact—self-management	55.3%	57.0%	1.5	1.5
Program belonging and engagement	53.6%	55.4%	1.4	1.5

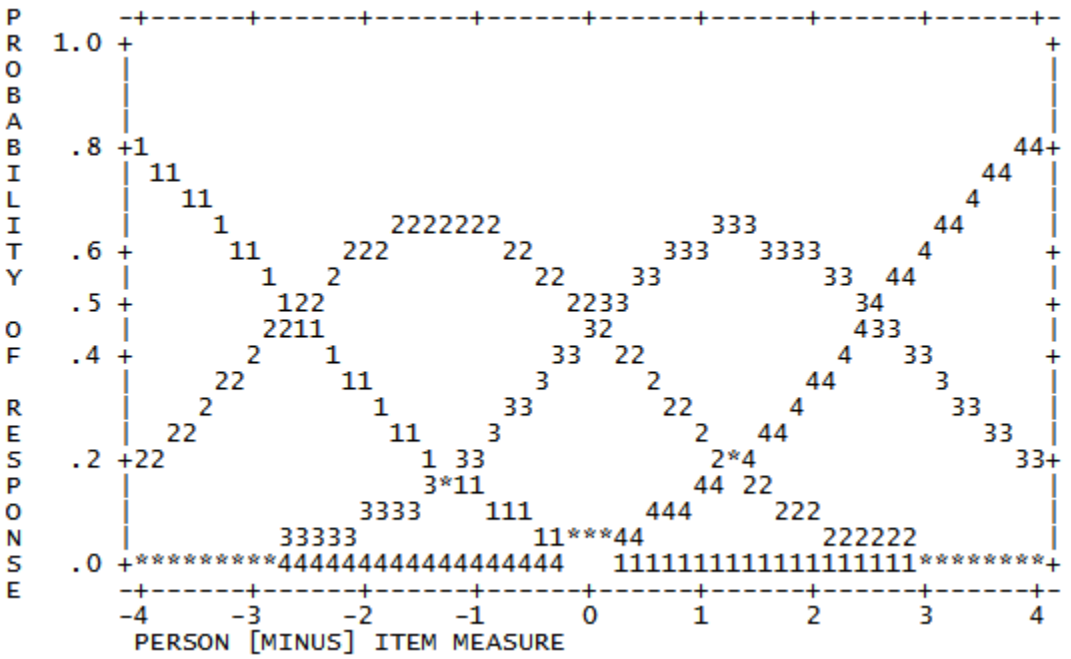
Are Survey Scales Functioning Well Psychometrically? Substantive Validity

Application of the Rasch rating scale model yields information about how well the rating scale associated with a given measure is functioning from a psychometric perspective, including information about the actual width of each response option associated with a given rating scale relative to the construct being measured. As noted previously, the rating scale used on the *Youth Motivation, Engagement, and Beliefs Survey* uses a four-point response system—"not at all true", "somewhat true", "mostly true", *or* "completely true". Typically, ordinal response options akin to those used on the survey are treated as covering an equal spectrum of the underlying construct of interest (picture a ruler divided into four equal segments, with the width of equal segment representing the portion of the ruler associated with a rating of "not at all true", "somewhat true", etc.).

When conducting Rasch analyses of this kind, however, the actual width of a response category is empirically based on how youth used the rating scale for the bank of items represented on a given scale. The point where one rating option transitions to another is called a step calibration. For example, Figure 1 outlines a probability curve pertaining to the rating scale structure of the *academic identity* scale from the 2015 sample. The *x*-axis of the figure represents the continuum

of youth functioning on the academic identity scale (again, the ruler metaphor is useful to conceptualize what this represents). The *y*-axis represents the probability that a youth of a given ability would endorse a particular rating option on an item appearing on the survey with an average item difficulty. In this sense, a youth with an ability estimate of –1 logits, for example, would be most likely to provide a response of "somewhat true (2)" to an item of average difficulty, and a youth with a logit ability of 1 would be most likely to provide a response of "mostly true (3)" to a similar item.

Figure 1. Example Probability Curve Associated with the Academic Identity Scale From the 2015 Sample



As shown in Figure 1, the transition point between "not at all true (1)" and "somewhat true (2)" occurs at –2.61 logits, while the transition point between "somewhat true (2)" and "mostly true (3)" occurs at .07 logits. This means that the portion of the latent construct covered by the "somewhat true" portion of the scale ranges from –2.61 and .07 logits, covering 2.68 logits. According to Linacre (2004), the recommended minimum advance between step calibrations for a four-category scale is 1.2 logits. Spans less than that suggest that respondents are having some difficulty distinguishing between response categories.

Across all scales appearing on the *Youth Motivation, Engagement, and Beliefs Survey* and in both samples, each rating scale was found to meet this criterion in terms of exceeding the minimum advance between step calibrations, indicating the four-point rating scale was functioning well for all components of the survey.

In addition to examining the extent to which step calibrations advance in a manner that conveys effective rating scale functioning, additional guidelines for assessing the quality of the rating scale associated with a given measure include the following:

1. Each rating scale category contains a minimum of 10 responses. This criterion was easily met with the sample size associated with each of the two samples. As we will see in later sections of the report, however, both the "not at all true (1)" and "somewhat true" categories were less frequently used, accounting for 2 to 11 percent of responses in the case of the former across the scales appearing on the survey and 13 to 22 percent of responses in the case of the latter.
2. Average respondent measure for each category increases monotonically. The average respondent measure is basically the average Rasch-calibrated score of the youth who responded with a given response option. One would expect that youth providing a higher response to a given item (for example, selecting "completely true" as opposed to "mostly true") would be more apt to have a higher score on a survey scale. This criterion also was found to be met for all scales across both samples.
3. Unweighted mean-squared fit indices within acceptable range. As with the assessment of item fit, fit indices also are calculated for each response option, with the range of 0.5 to 1.5 used to evaluate good fit. All response categories across each scale in both samples were found to be within this range.
4. Smooth and unimodal shape to each rating scale. This criterion is shown by the probability curve shown in Figure 1, where each rating scale option is shown to peak for a portion of scale before descending. The probability curve calibrated for each scale and sample was similar to the one appearing in Figure 1, indicating each rating scale met this criterion.

In light of these findings, the rating scale used on the revised version of the *Youth Motivation, Engagement, and Beliefs Survey* was found to functioning quite well from a psychometric perspective, with no major need for modification or revision.

Are Survey Scales Functioning Well Psychometrically? Generalizability/Reliability

One of the primary objectives associated with most efforts to validate a measure like the *Youth Motivation, Engagement, and Beliefs Survey* concerns assessing the reliability of the measures resulting from use of the tool in question. Most commonly, reliability is assessed by calculating the Cronbach's alpha statistic. Cronbach's alpha is a measure of the internal consistency of a scale, describing the extent to which all of the items in the scale measure the same concept. The higher the alpha, the more highly correlated the items are with one another. As shown in Table 6, alphas ranged from 0.81 to 0.92 for the full domain of scales appearing on the survey across both samples. Generally, alphas above 0.80 are indicative of good scale reliability.

Table 6. Cronbach's Alpha and Person Separation Reliability by Subscale

	Cronbach's Alpha		Person Separation Reliability	
Subscale	**2015**	**2016**	**2015**	**2016**
Academic identity	0.87	0.89	0.65	0.69
Mindsets	0.83	0.84	0.71	0.74
Self-management	0.81	0.82	0.71	0.73
Interpersonal skills	0.81	0.84	0.68	0.71
Program impact—academics	0.90	0.90	0.75	0.78
Program impact—self-management	0.91	0.92	0.81	0.83
Program belonging and engagement	0.92	0.92	0.75	0.78

In addition, scale calibrations in Rasch also produce what is termed a person separation reliability index, which is a measure of how well the scale can distinguish among individuals performing at different levels on the construct of interest. For example, a measure with high person reliability is able to accurately place a person with lower ability or functioning on the lower end of the scale and a person with higher ability on the higher end of the scale. As with Cronbach's alpha, a person separation reliability index above 0.80 is indicative of good measure function, values between 0.70 and 0.80 would be deemed acceptable, and values between 0.60 and 0.70 questionable.

As shown in Table 6, for most scales appearing on the survey, the person reliability index was between 0.70 and 0.80 for both samples, indicating the scales were functioning at an acceptable level in this regard. The primary exception to this finding was the academic identity scale, where the person separation index was 0.65 and 0.69 for the 2015 and 2016 samples respectively. It is important to note that the academic identity scale was the shortest scale, composed of only five items. Adding items to the scale so it is equivalent in length to the other scales (that largely are composed of six to nine items) would likely improve person reparation reliability estimates.

In addition, for the 2015 sample, the person separation reliability index also was below 0.70 for the interpersonal skills scale, although for the 2016 sample, the index was found to exceed the 0.70 threshold. This again is a smaller scale, with only six items. The addition of one well-constructed item would likely move this scale to consisting performing upon 0.70 on the person separation reliability index.

Are Survey Scales Functioning Well Psychometrically? External Validity

Responsiveness/Sensitivity

Another important facet to assess in determining how well a measure is functioning is whether floor or ceiling effects are associated with any subscale and as a consequence, the degree to which the measure is likely to be sensitive to detecting change over time. It is common for surveys like *Youth Motivation, Engagement, and Beliefs Survey* to be characterized by ceiling effects, meaning a significant percentage of respondents end up with the maximum score on a scale. This would occur, for example, when a youth answered "completely true" to all items

appearing on the academic identity scale. There are two primary consequences when ceiling effects are found to characterize a measures: (1) the estimate of the true level of functioning of the respondent is unknown because there was an insufficient number of items on the scale that were harder for the respondent to agree with, and as a result, the respondent's true level functioning lies somewhere beyond what the current set of items can assess and (2) it makes the measures less useful as a mechanism to document change over time if a significant number of respondents have no room to grow on the scale in question. Floor effects have the reverse effect, in which the items are too difficult for respondent's to agree with, and as a result, respondents' true level of functioning lies somewhere below what the current items measure. Fortunately, no floor effects were found to characterize the domain of scales appearing on the *Youth Motivation, Engagement, and Beliefs Survey*, but the same cannot be said in relation to ceiling effects.

In Table 7, the percentage of respondents in both the 2015 and 2016 samples who received the maximum score on each scale is outlined. In reviewing these results, a couple of distinctions need to be made. The presence of ceiling effects is of particular concern in relation to those scales that are meant to assess youth reported functioning when taking the survey on a series of areas related to positive youth development—academic identity, positive mindsets, self-management, and interpersonal skills. These scales represent the types of youth outcomes 21st CCLC programs are trying to engender and represent the portion of the survey where there is a desire to meaningfully support youth growth and development of these skills, beliefs, and attitudes over time. The other scales appearing on the survey are meant solely to capture cross-sectional information associated with a particular time point. In this regard, we are less concerned with ceiling effects observed in relation to the program impact—academics, program impact—self-management, and program belonging and engagement scales.

As shown in Table 7, between 5 and 9 percent of youth across both sample received the maximum score possible on the positive mindsets, self-management, and interpersonal skills scales, suggesting that ceiling effects are not an overwhelming issue on these scales. The percentage of youth receiving the maximum score in relation to the academic identity scale, however, was substantially higher, with more than 20 percent of youth falling in this category in both samples. This finding is another indication that this scale warrants some revision, particularly by adding items that are harder for respondents to endorse to address the ceiling effects observed in Table 7 and improve person separation reliability.

Table 7. Percentage of Respondents with Maximum Scale Score by Sample

	Percentage of Youth With Maximum Scale Score	
Subscale	**2015**	**2016**
Academic identity	22.1%	23.2%
Mindsets	6.1%	6.4%
Self-management	5.1%	5.2%
Interpersonal skills	8.1%	8.6%
Program impact—academics	20.2%	18.4%
Program impact—self-management	12.5%	11.9%
Program belonging and engagement	20.0%	18.2%

Scale scores resulting from the application of Rasch approaches also can be used to classify the portion of the rating scale in which a youth's survey responses fell. This provides us with another opportunity to explore how youth were classified across each of the scales represented on the survey and what this means for potentially assessing the current level of youth functioning and the potentiality for growth and development on survey scales over time.

In Figure 2, steps have been taken to outline the distribution of youth across each of the survey response categories for those survey scales related to the types of youth development outcomes programs are trying to engender. As shown in Figure 2, anywhere from 78 to 90 percent of responding youth were classified in either the "mostly true" or the "completely true" portion of the rating scale, with nearly 50 percent of youth falling in the "mostly true" category for *positive mindsets, self-management,* and *interpersonal skills*. Here again, we see evidence of a ceiling effect in relation to the *academic identity* scale, with more than 50 percent of responding youth falling in the "completely true" category. The scale demonstrating the most opportunity for growth is the *self-management* scale, where more than 20 percent of respondents fell in the "not at all true" or "somewhat true" categories.

The distribution of youth responses across scales related to self-reported program impact and feelings of program belonging engagement are shown in Figure 3. As outlined in Figure 3, there is substantially more variation across response categories for both the *program impact—academics (RAB)* and *program impact—self-management (RSM)* scales than what was observed in relation to the scales outlined in Figure 2, although the majority of responses still fell in the "mostly true" and "completely true" portions of the scale. Generally, youth were slightly more likely to report program impact in relation to academics than in relation to the development of self-management skills. In terms of the *program belonging and engagement* scale, in excess of 80 percent of youth in both samples indicated that items describing a positive experience in programming as falling in the "mostly true" and "completely true" portions of the scale

Figure 2. Distribution of Youth Across Rating Scale Categories by Scale and Sample Year: Scales Related to Youth Outcomes

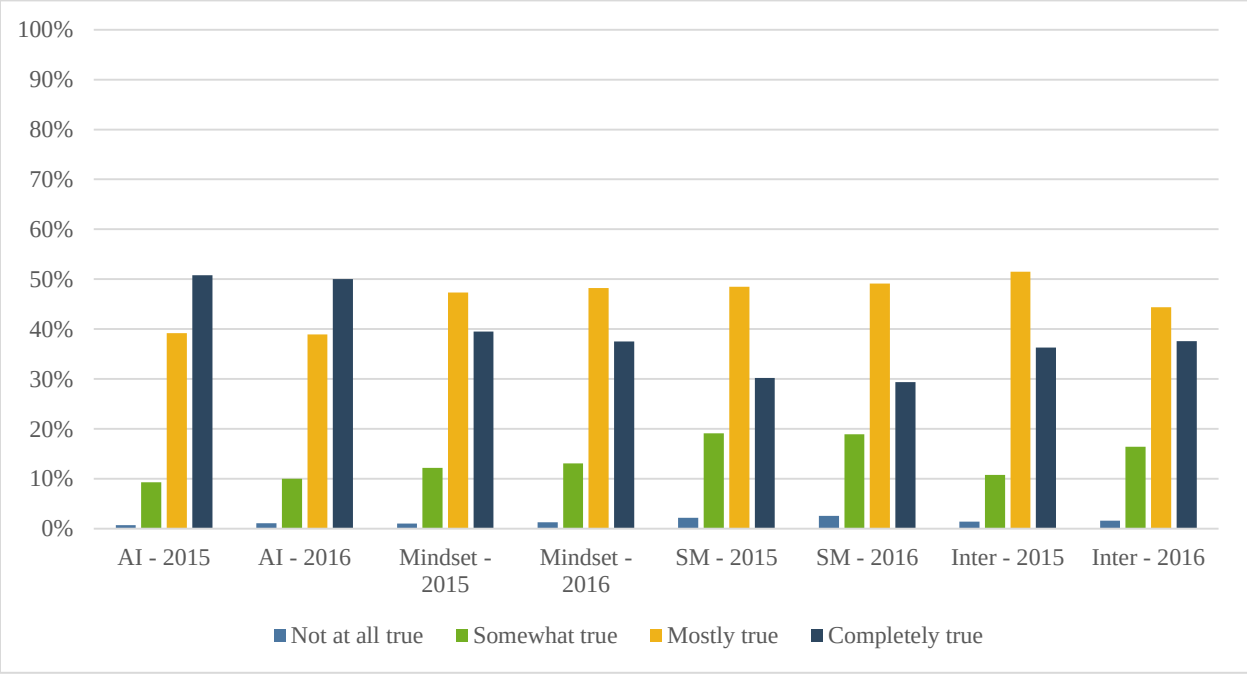
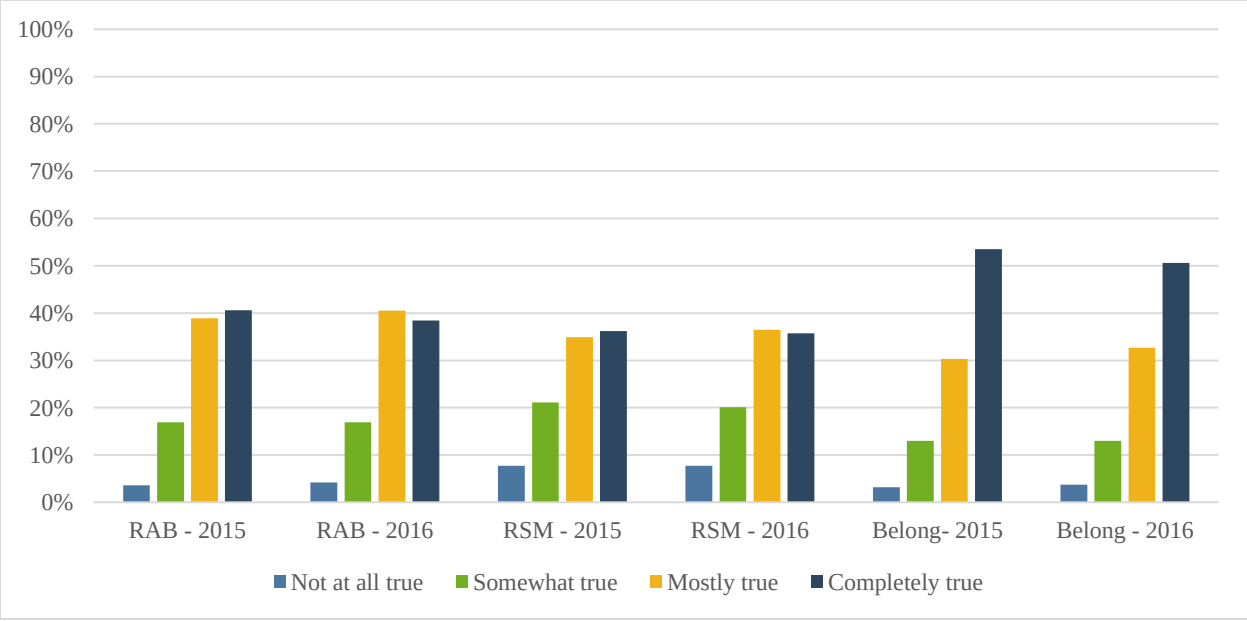


Figure 3. Distribution of Youth Across Rating Scale Categories by Scale and Sample Year: Scales Related to Perceived Program Impact and Program Experiences



As mentioned previously, one of the potential purposes of the *Youth Motivation, Engagement, and Beliefs Survey* is to measure growth on the domain of youth outcomes measured on the survey; as shown in Figure 2, however, because of the extremely high percentage of youth who fell within the "mostly true" and "completely true" portions of the scale, the viability of using the survey for this

purpose could be called into question. In order to explore this issue further, two additional types of analyses were done—(1) the calculation of a person strata index and (2) the comparison of pre-post data from youth taking the survey both in spring of 2015 and in spring of 2016.

The person strata index indicates the number of levels of a trait like *positive mindsets* that a measure is able to distinguish from the items making up that measure (Wolfe & Smith, 2007). For example, as Wolfe and Smith note, if you trying to use a measure to compare how individuals enrolled in an experimental and control group were doing on some trait where you expected the experimental group to do better because of the intervention they received, you would want the measure to be able to distinguish between at least two levels of the trait in question. As shown in Table 8, across both samples, most of the measures related to youth outcomes are able to distinguish between at three levels of the trait in question, and the *positive mindset* and *self-management* scales allow for four levels to be distinguished in the 2016 sample. These results seem to support the viability of using the *Youth Motivation, Engagement, and Beliefs Survey* as a pre-post measure, although the utility of doing so also relates to the number of youth falling in each category that can be distinguished by the measure. As shown in Figure 2, roughly 30 percent to 40 percent of youth, and in the case of *academic identity*, 50 percent of responding youth, were found to fall in the "completely true" portion of the scale. This leaves approximately 50 percent to 70 percent of youth in a position where some substantial growth on the scale is possible, although the bulk of these youth are currently falling within the "mostly true" portion of the scale.

Table 8. Person Strata Index for the Outcomes Scales by Sample

	Person Strata Index	
Subscale	**2015**	**2016**
Academic identity	2.81	3.30
Mindsets	3.60	4.13
Self-management	3.60	3.94
Interpersonal skills	3.17	3.60

In order to further assess how sensitive survey scales associated with youth outcomes were to detecting changes in youth functioning on the constructs of interest appearing on the survey, steps were taken to identify a subset of youth that completed the survey in both spring 2015 and spring 2016. A total of 984, or 22 percent of youth in the 2015 sample, were found to have taken the survey in both years.

As shown in Table 9, the overall mean scores for youth taking in the survey in both 2015 and 2016 declined slightly from time 1 to time 2,[1] and although these declines were found to be significant for three of the four scales in question based on a paired sample *t*-test, the degree of decline was for all practical purposes was close to 0, with the large sample size driving the

[1] In calculating these means, the logit value resulting from Rasch calibrations were converted to a 1 through 4 scale to better represent the four-point response scale associated with the survey.

significant results. In addition, the correlation between 2015 and 2016 scores was found to be moderate for each scale, ranging from .327 to .378.

Table 9. Comparison of 2015 and 2016 Scores by Subscale—Full Sample

	Paired Sample *t*-Test			Bivariate Correlation	
Subscale	**2015 Mean**	**2016 Mean**	***p* value**	**Correlation Coefficient**	***p* value**
Academic identity (*n* = 982)	3.31	3.25	.004**	.378	.000***
Mindsets (*n* = 981)	3.00	2.94	.002**	.334	.000***
Self-management (*n* = 981)	2.82	2.81	.209	.304	.000***
Interpersonal skills (*n* = 981)	3.01	2.97	.047*	.327	.000***

***$p < .001$,**$p < .01$, *$p < .05$

Next, steps were taken to explore how changes in survey scores might be different for youth that (1) fell in the "not at all true" and "somewhat true" portions of the scale in spring 2015 and (2) youth receiving a scale score in the bottom 50th percentile for the scale in question. These results are shown in Tables 10 and 11 respectively. As shown in Table 10, youth scoring in the bottom two response categories of the survey demonstrated substantive growth between the 2015 and 2016 administrations, ranging from .37 to .56 scale score points on the four-point scale. As might be expected given the level of change, 2015 and 2016 scores for this group were weakly and not significantly correlated.

A similar trend was found in Table 11, where youth falling in the bottom 50th percentile of each scale based on their 2015 survey were considered. Here, improvements ranged from .16 to .25 scale score points, and all correlations between 2015 and 2016 scores, were found to be moderately and significantly correlated.

Table 10. Comparison of 2015 and 2016 Scores by Subscale—Bottom Two Response Categories

	Paired Sample *t*-Test			Bivariate Correlation	
Subscale	**2015 Mean**	**2016 Mean**	***p* value**	**Correlation Coefficient**	***p* value**
Academic identity (*n* = 83)	2.19	2.75	.000***	.093	.401
Mindsets (*n* = 106)	2.27	2.64	.000***	.026	.790
Self-management (*n* = 195)	2.21	2.59	.000***	.115	.109
Interpersonal skills (*n* = 104)	2.20	2.66	.000***	.095	.337

***$p < .001$

Table 11. Comparison of 2015 and 2016 Scores by Subscale—Bottom 50th Percentile

	Paired Sample *t*-Test			Bivariate Correlation	
Subscale	**2015 Mean**	**2016 Mean**	***p* value**	**Correlation Coefficient**	***p* value**
Academic Identity (*n* = 437)	2.78	3.03	.000***	.270	.000***
Mindsets (*n* = 457)	2.63	2.79	.000***	.213	.000***
Self-Management (*n* = 491)	2.47	2.68	.000***	.197	.000***
Interpersonal Skills (*n* = 477)	2.60	2.82	.000***	.236	.000***

****p* < .001

The results from Tables 9 to 11 seems to suggest the following conclusions on the utility of the *Youth Motivation, Engagement, and Beliefs Survey* to assess changes in youth functioning over time. First, the mean scores for the sample with both 2015 and 2016 scores were quite stable, demonstrating a very slight decline between the two administration periods, although pre and post scores were found to be only moderately correlated. When there was room for youth to grow on the scales in question, however, significant and substantive growth was shown for youth scoring both in the bottom two response categories in spring 2015 and in the bottom 50th percentile of a given scale. A couple of preliminary hypotheses can be made about the nature of this positive growth for these populations. One, this growth could actually represent growth that took place during this period and it could be that the participation in 21st CCLC may have contributed to this growth. Unfortunately, we do not have the data to rigorously explore whether this was the case at this point in time. In addition, it also could be that youth with lower levels of functioning in spring 2015 simply regressed back to the mean of the overall sample between administrations, and the survey is really not capturing any real growth between the two time periods.

One approach that can be taken to explore which of these explanations may be more viable is to explore how growth on the survey between the two administration periods may be related to other data collected on the *Youth Motivation, Engagement, and Beliefs Survey*. For example, we can explore the question "How is growth on survey subscales between 2015 and 2016 related to the experiences youth had in programming based on responses to the program belonging and engagement scale?" Our hypothesis would be that better program experiences would be associated with larger growth on survey scales. And so, steps were taken to assess this hypothesis for each of the samples outlined in Tables 9 to 11 by examining the correlation between pre-post scale changes and the youth's scale score on the *program belonging and engagement* scale for the survey taken in spring 2016. The results of these analyses are shown in Table 12. Not only was the degree of growth positively and significantly correlated with the program belonging and engagement scale, but also the strength of the correlation increased as the sample of youth increasingly represented youth who scored lowest on survey scales during the

2015 administration, with the correlation ranging from .587 to .651 for youth who scored in the bottom two response categories ("not at all true" and "somewhat true") in 2015 for a given scale. This type of result would seem to support a possible connection between program experiences and growth on survey scales among youth demonstrating a lower level of functioning in these areas at baseline.

Similar results were found when the degree of growth in the academic identity scale was correlated with the scale scores for the *program impact—academic* scale and growth on the *self-management* scale was correlated with the scale scores for the *program impact—self-management* scale as shown in Tables 13 and 14. In this sense, for youth demonstrating a lower level of functioning on these scales at baseline, the correlation between growth on these scales and self-reported impact was stronger than for youth performing at higher level on these scales in spring 2015.

Table 12. Correlation of Changes in Scale Score With Responses to the Program Belonging and Engagement Scale by Survey Sample

	Full Sample		Bottom 50th Percentile		Bottom Two Response Categories	
Subscale	**Correlation Coefficient**	***p* value**	**Correlation Coefficient**	***p* value**	**Correlation Coefficient**	***p* value**
Academic identity	.321	.000***	.465	.000***	.601	.000***
Mindsets	.374	.000***	.526	.000***	.651	.000***
Self-management	.363	.000***	.481	.000***	.587	.000***
Interpersonal skills	.381	.000***	.554	.000***	.601	.000***

****p* < .001

Table 13. Correlation of Changes in Academic Identity Scale Score \With Responses to the Program Impact—Academics Scale by Survey Sample

	Full Sample		Bottom 50th Percentile		Bottom Two Response Categories	
Subscale	**Correlation Coefficient**	***p* value**	**Correlation Coefficient**	***p* value**	**Correlation Coefficient**	***p* value**
Academic identity	.304	.000***	.474	.000***	.663	.000***

****p* < .001

Table 14. Correlation of Changes in Self-Management Scale Score With Responses to the Program Impact—Self-Management Scale by Survey Sample

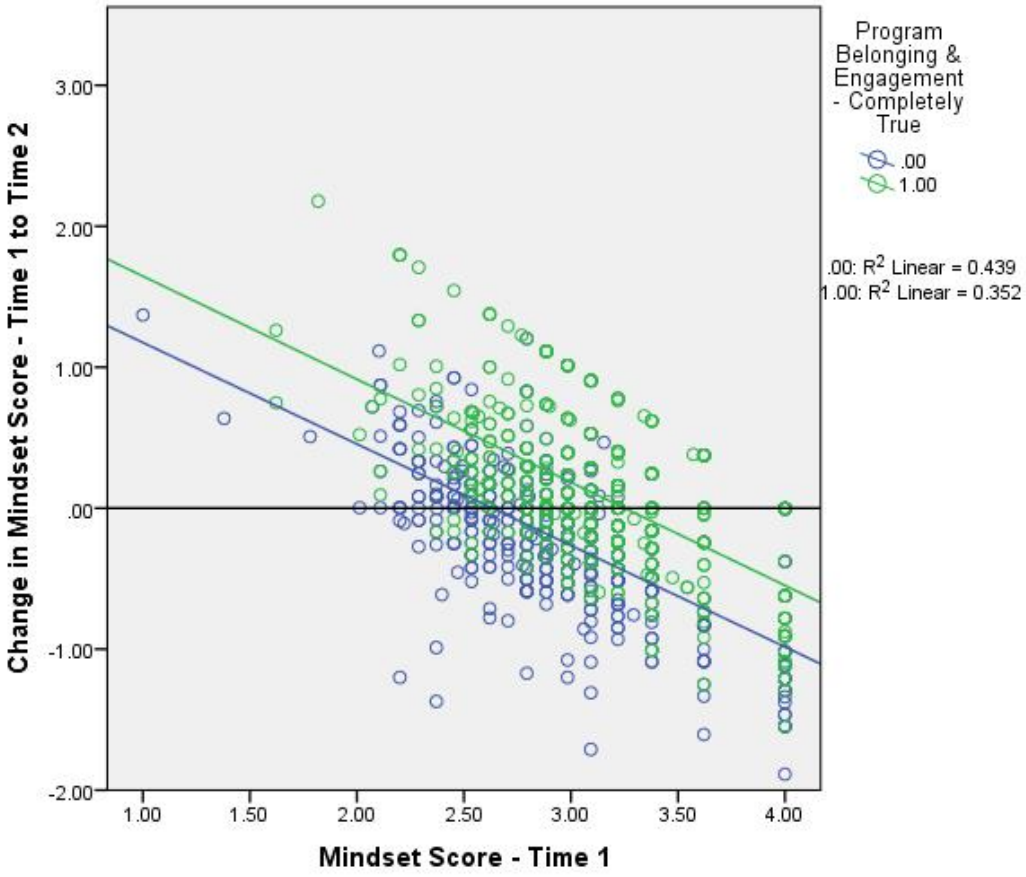
	Full Sample		Bottom 50th Percentile		Bottom Two Response Categories	
Subscale	**Correlation Coefficient**	***p* value**	**Correlation Coefficient**	***p* value**	**Correlation Coefficient**	***p* value**
Self-management	.346	.000***	.489	.000***	.613	.000***

****p* < .001

Another approach that can be used to explore the possibility that pre-post changes are associated with the regression to the mean are to examine scatterplots of the scores received by youth at time 1 and the level of growth witnessed between times 1 and 2. An example of this is shown in Figure 4 in relation to the *mindsets* scale. As shown in Figure 4, growth on the *mindset* scale is shown on the *y*-axis. The solid line at 0 indicates no growth between pre administration and post administration. Circles above the 0 line indicate cases where youth demonstrate improvement between time 1 and time 2. Circles below the 0 line demonstrate a decrease. As shown in the scatterplot, youth scoring lower on the mindset scale at time 1 were more apt to show improvement while youth scoring higher at time 1 were more likely to show a decline. This type of pattern is indicative of the presence of regression to the mean.

In addition, in Figure 4, two regression lines have been added to the chart, one for youth that scored in the "completely true" range of the *program belonging and engagement* scale (the higher line) and one for youth that scored in the other response categories making up the scale in question (that is, "not at all true," "somewhat true," and "mostly true"—the lower line in Figure 4). The distance between the lines represents the relationship described previously that more positive experiences in the program are associated with more growth on the mindset scale between time 1 and time 2. Scatterplots created for *academic identity*, *self-management*, and *interpersonal skills* demonstrated similar results to those outlined in Figure 4.

Figure 4. Scatterplot of Time 1 Score on the Mindsets Scale and Change from Time 1 to Time 2



In this sense, Figure 4 suggests there are likely issues related to regression to the mean between pre administration and post administrations of the survey that should be controlled for when examining pre and post change using the survey, although again survey scales do appear to be potentially sensitive to capturing the hypothesized relationship between positive program experiences and growth on youth outcomes.

Is the Level of Youth Functioning on Survey Scales Predictive of School-Related Outcomes in the Manner Hypothesized? Predictive Validity

The constructs measured in the *Youth Motivation, Engagement, and Beliefs Survey* were the result of extensive efforts undertaken by YDEKC to identify and measure key skills and beliefs related to positive youth growth and development. In light of this, we wanted to explore whether youth functioning on survey scales would be related to a series of school-related outcomes obtained from the data warehouses maintained by OPSI. The hypothesis here is that higher scale scores would be found to be related to a variety of positive school-related outcomes, thereby empirically demonstrating the potential connection between what is measured on the survey and the types of academic-related outcomes sought by the 21st CCLC program.

When collecting youth survey data, steps also were taken to capture the unique statewide identifier for each youth, allowing survey response data to be linked to school-related demographic and outcome data housed in OSPI's data warehouse. Of the 4,497 youth completing the survey from the 2015 sample, matches were found for 3,463 youth in Grades 4–8 (a relatively small number of youth in Grades 9–12 were represented in the sample and were therefore excluded from analyses described in this section of the report as well).

In order to explore this possible relationship, a series of HLMs models were run to assess the correlation between survey scale scores and school-related outcomes associated with the 2015 school year. The following outcome variables were included in these analyses:

- State assessment scores in reading
- State assessment scores in mathematics
- Number of absences
- Number of disciplinary incidents
- Number of intervention days associated with disciplinary incidents

A series of other youth- and school-level predictors were included in the model to control for key features related to the school-related outcomes in question:

Youth-level:

- Eligibility for free and reduced-price lunches
- Special education status
- Bilingual status
- Hispanic ethnicity
- Enrollment in the Learning Assistance Program for reading
- Enrollment in the Learning Assistance Program for mathematics

School-level:

- Number of youth enrolled in the school
- Percentage of school population that is Hispanic
- Percentage of school population eligible for free and reduced-price lunches
- Percentage of school population with bilingual status
- Percentage of school population with special education status
- Mean number of unexcused absences
- Ratio of the number of disciplinary incidents at the school to school enrollment

Separate models were run for each survey subscale and outcome, where a youth's score on a given scale was included in the model as a level-one predictor. The goal here was to examine whether a given subscale was found to be related to a given school-related outcome.

As outlined in Table 15, higher scores on the *academic identity* scale were found to be significantly related to higher reading and mathematics assessment scores, fewer unexcused absences, fewer disciplinary incidents, and fewer intervention days.

Higher scores on the *mindset* scale also were found to be significantly related to higher reading and mathematics assessment scores, fewer unexcused absences, fewer disciplinary incidents, and fewer intervention days.

Higher scores on the *self-management* scale were found to be significantly related to higher mathematics assessment scores, fewer unexcused absences, fewer disciplinary incidents, and fewer intervention days. There was also a moderately significant relationship with higher reading assessment scores.

Higher scores on the *interpersonal skills* scale also were related to higher reading assessment scores, fewer unexcused absences, fewer disciplinary incidents, and fewer intervention days. Higher scores on the interpersonal skills scale also were associated with higher mathematics assessment scores, although this relationship was not statistically significant.

These results further lend support to case that the scales appearing on the *Youth Motivation, Engagement, and Beliefs Survey* are measuring survey constructs in a way that have been shown to be related to school-related outcomes in the manner predicted. AIR plans to replicate these analyses in 2016 for the 2015–16 programming period to see whether these result continue to hold up and to assess the stability and sensitivity of scores across time to see whether the tool can be viably be used to measure youth growth on these constructs over time.

Table 15. Summary of HLM Results by Survey Subscale and School Outcome

	Coefficient	Standard Error	*p* value
Academic identity			
Reading assessment	0.060	0.008	0.000***
Mathematics assessment	0.059	0.007	0.000***
Unexcused absences	-0.092	0.006	0.000***
Disciplinary incidents	-0.183	0.018	0.000***
Intervention days	-0.211	0.016	0.000***
Mindsets			
Reading assessment	0.034	0.010	0.000***
Mathematics assessment	0.045	0.010	0.000***
Unexcused absences	-0.078	0.008	0.000***
Disciplinary incidents	-0.146	0.024	0.000***
Intervention days	-0.122	0.022	0.000***
Self-management			
Reading assessment	0.019	0.011	0.078†
Mathematics assessment	0.022	0.011	0.040*
Unexcused absences	-0.090	0.009	0.000***
Disciplinary incidents	-0.225	0.028	0.000***
Intervention days	-0.249	0.025	0.000***
Interpersonal skills			
Reading assessment	0.018	0.009	0.054*
Mathematics assessment	0.001	0.009	0.943
Unexcused absences	-0.046	0.008	0.000***
Disciplinary incidents	-0.200	0.025	0.000***
Intervention days	-0.151	0.022	0.000***

N = 3,463 youth in grades 4–8 with complete survey data, actual sample size varies by analysis

***$p < .001$, **$p < .01$, *$p < .05$, †$p < .10$

Summary and Conclusions

The purpose of this report was to summarize a series of analyses conducted to assess the reliability and validity of the *Youth Motivation, Engagement, and Beliefs Survey* based on data collected by AIR in relation to the survey as part of the statewide evaluation of the Washington 21st CCLC program. More specifically, the results outlined in this report were oriented at answering two primary questions.

1. Are survey scales functioning well psychometrically?
2. Is the level of youth functioning on survey scales predictive of school-related outcomes in the manner hypothesized?

Overall, the answer to the first question is primarily affirmative. On most criteria examined in relation to the psychometric functioning of the survey, survey scales were by and large found to function quite well in terms of item fit, rating scale functioning, unidimensionality, reliability, and sensitivity to change That being said, some modifications may be warranted, and some issues may require additional study.

For example, the Rasch person separation reliability coefficients (which is a measure of how well the scale can distinguish among individuals performing at different levels on the construct of interest) for the *academic identity* scale were found to be lower than desired for both the 2015 and 2016 samples, and given the ceiling effects found to be associated with this scale, it seems that scale functioning could be improved by adding two well-written items that are more difficult for youth being able to agree with. A similar issues was found to characterize the *interpersonal skills* scale, which also could benefit for an additional item to improve person separation reliability.

In addition, for the four scales oriented at measuring youth functioning in four key outcome areas (that is, *academic identity, positive mindsets, self-management,* and *interpersonal skills*), youth had a tendency to be classified as falling in the "mostly true" and "completely true" portions of the rating scale. This certainly has ramifications for being able to use the *Motivation, Engagement, and Beliefs Survey* to measure growth on a pre to post basis for the full domain of youth served in afterschool or youth development program, although results suggest the potential viability of the tool for assessing improvement in these areas for youth scoring in the bottom 50th percentile on a given scale or that smaller subset of youth falling in either the "not at all true" or "somewhat true" portions of the scale.

Some evidence suggests that regression to mean may account for some of the growth shown on each of the survey scales. Any effort to use the survey to assess pre to post change needs to include steps to control for this issue through either the research design and/or statistical adjustment.

Although we were able to find a relationship between improvement on survey scales and youth experiences in programming and self-reported impact, our sense is that additional work needs to

be done in this area to assess how growth on survey scales may be related to other desirable outcomes, including school-related outcomes. Part of these efforts should also explore whether growth on survey scales demonstrated by youth already falling in the "mostly true" portion of the scale has substantive and meaningful ramifications for how they may function in other settings like school. Some of this work be undertaken by AIR as part of evaluation activities associated with the 2016–17 21st CCLC programming period.

Finally, efforts to explore if survey scales were predictive of school-related outcomes found significant relationships between survey scale scores and school-related outcomes in the manner hypothesized across all scales and outcomes examined. This is a promising finding and suggests that what is being measured on the *Youth Motivation, Engagement, and Beliefs Survey* has relevance to youth functioning in other contexts as hypothesized by the tool developers at YDEKC.

Our conclusion based on the domain of results summarized in this report is that the *Youth Motivation, Engagement, and Beliefs Survey* is a promising tool for measuring many important elements of youth functioning that afterschool and youth development programs are seeking to cultivate and that are important to youth success in school and life more broadly.

References

Devaney, E. (2015a). *Social and emotional learning practices: A self-reflection tool for afterschool staff.* Washington, DC: American Institutes for Research.

Devaney, E. (2015b). *Supporting social and emotional development through quality afterschool programs.* Washington, DC: American Institutes for Research.

Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of afterschool programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology*, 45, 294–309.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance--A critical literature review*. Chicago, IL: Consortium on Chicago School Research.

Larson, R. W., & Angus, R. M. (2011). Adolescents' development of skills for agency in youth programs: Learning to think strategically. *Child Development*, 82, 277–294.

Larson, R. W., & Dawes, N. P. (In press). How to cultivate adolescents' motivation: Effective strategies employed by the professional staff of American youth programs. In Stephen Joseph (Ed.), *Positive psychology in practice*. New York: Wiley.

Moroney, D. (2016). *The readiness of the out-of-school time workforce to intentionally support participants' social and emotional development: A review of the literature and future directions*. Keynote address for the Workshop on Character Education for the National Academies of Sciences, July 26, 2016, Washington, D.C.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr., & R. M. Smith (Eds.), *Introduction to Rasch measurement (*pp. 258–278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2009). *Winsteps, version 3.71.0* [Computer software].

Smith, C., McGovern, G., Larson, R., Hillaker, B., & Peck, S. C. (2016). *Preparing youth to thrive: Promising practices for social and emotional learning*. Washington, DC: Forum for Youth Investment.

Wilson-Ahlstrom, A., Yohalem, N., DuBois, D. L., Ji, P., & Hillaker, B. (2014). *From soft skills to hard data: Measuring youth program outcomes*. Washington, DC: Forum for Youth Investment.

Wolfe, E. W., & Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. In E. V. Smith Jr., & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243–290). Maple Grove, MN: JAM Press.

Appendix A. Youth Motivation, Engagement, and Beliefs Survey

	Not at all true	*Somewhat true*	*Mostly true*	*Completely True*
ACADEMIC IDENTITY				
Doing well in school is an important part of who I am.	1	2	3	4
Getting good grades is one of my main goals.	1	2	3	4
I take pride in doing my best in school.	1	2	3	4
I am a hard worker when it comes to my schoolwork.	1	2	3	4
It is important to me to learn as much as I can.	1	2	3	4
MINDSETS				
I finish whatever I begin.	1	2	3	4
I stay positive when things don't go the way I want.	1	2	3	4
I don't give up easily.	1	2	3	4
I try things even if I might fail.	1	2	3	4
I can solve difficult problems if I try hard enough.	1	2	3	4
I can do a good job if I try hard enough.	1	2	3	4
I stay focused on my work even when it's boring.	1	2	3	4
SELF-MANAGEMENT				
I can stop myself from doing something I know I shouldn't do.	1	2	3	4
When I'm sad, I do something that will make me feel better.	1	2	3	4
I can control my temper.	1	2	3	4
I can handle stress.	1	2	3	4
I can calm myself down when I'm excited or upset.	1	2	3	4
When my solution to a problem is not working, I try to find a new solution.	1	2	3	4
I think of my past choices when making new decisions.	1	2	3	4
INTERPERSONAL SKILLS				
I listen to other people's ideas.	1	2	3	4
I work well with others on group projects.	1	2	3	4
I feel bad when someone gets their feelings hurt.	1	2	3	4
I respect what other people think, even if I disagree.	1	2	3	4
I try to help when I see someone having a problem.	1	2	3	4
When I make a decision, I think about how it will affect other people.	1	2	3	4

Program Experiences Scales

	Not at all true	*Somewhat true*	*Mostly true*	*Completely True*
Academic Behaviors (retrospective)				
This program has helped me to become more interested in what I'm learning in school	1	2	3	4
This program has helped me to connect my schoolwork to my future goals	1	2	3	4
This program has helped me to do better in school	1	2	3	4
This program has helped me to complete my schoolwork on time	1	2	3	4
This program has helped me to do a better job on my schoolwork	1	2	3	4
Self-Management (retrospective)				
This program has helped me to handle stress	1	2	3	4
This program has helped me to become better at controlling my temper	1	2	3	4
This program has helped me learn that my feelings affect how I do at school	1	2	3	4
This program has helped me learn how to be patient with others	1	2	3	4
This program has helped me learn how to calm myself down when I'm excited or upset	1	2	3	4
This program has helped me get better at staying focused on my work	1	2	3	4
This program has helped me learn to resist doing something when I know I shouldn't do it	1	2	3	4
BELONGING AND ENGAGEMENT SCALE				
I feel proud to be part of my program	1	2	3	4
What we do in this program will help me succeed in life	1	2	3	4
There are things happening in this program that I feel excited about	1	2	3	4
This program helps me explore new ideas	1	2	3	4
This program helps me build new skills	1	2	3	4
What we do in this program is important to me	1	2	3	4
What we do in this program is challenging in a good way	1	2	3	4