



the  
**forum**  
FOR YOUTH INVESTMENT

Published by The Forum for Youth Investment  
January 2014

# Technical Appendix

FROM SOFT SKILLS TO HARD DATA:  
MEASURING YOUTH PROGRAM OUTCOMES

David DuBois, University of Illinois at Chicago,  
Peter Ji, Adler School of Professional Psychology &  
Barbara Hillaker, David P. Weikart Center for Youth Program Quality



# Table of Contents

Framework and Criteria for Ratings of Reliability and Validity Evidence { **3** }

Technical Summaries of Scales from the *Resilience & Youth Development Module* of the *California Healthy Kids Survey* { **7** }

Technical Summaries of Scales from the *Developmental Assets Profile* { **12** }

Technical Summaries of Scales from the *Devereux Student Strengths Assessment* { **15** }

Technical Summaries of Scales from the *Holistic Student Assessment* { **22** }

Technical Summaries of Scales from the *San Francisco Beacons Youth Survey* { **29** }

Technical Summaries of Scales from the *Social Skills Improvement System* { **34** }

Technical Summaries of Scales from the *Survey of Academic and Youth Outcomes* { **49** }

Technical Summaries of Scales from the *Youth Experiences Survey 2.0* { **63** }

Technical Summaries of Scales from the *Youth Outcomes Battery* { **73** }

Technical Summaries of Scales from the *Youth Outcome Measures Online Toolbox* { **79** }

# Technical Appendix

## From Soft Skills to Hard Data: Measuring Youth Program Outcomes

### Framework and Criteria for Ratings of Reliability and Validity Evidence

This Appendix provides an overview of the procedural steps and guidelines that were used in arriving at the ratings of reliability and validity evidence that are reported for each of the scales reviewed in this guide.<sup>1</sup> An overview of the framework used is shown below (Figure 1).<sup>2</sup> Those interested can obtain a copy of the complete rating system used from the authors of *From Soft Skills to Hard Data: Measuring Youth Program Outcomes* upon request. There are inherent limitations to any effort that is made to boil down the often varied and nuanced sources of evidence that bear on the psychometric properties of a measure into summative ratings. Users of this guide are encouraged to be mindful of this reality and to always consider the ratings that are provided for a scale in conjunction with the narrative accounts that are provided of the underlying evidence.

For each scale, the rating process began with the following set of general orienting questions:

- What construct is the measure intended to assess?
- For what types of youth populations (age, gender, ethnicity, etc.) is the measure intended to be appropriate?
- For what types of raters (youth, OST program staff, teacher, etc.) is the measure intended to be appropriate?

<sup>1</sup> The overall reliability and validity evidence for each of the instruments included in this guide was also evaluated. These assessments took into account both reliability and validity evidence for each of the different individual scales on an instrument. We also considered evidence for the validity of the instrument's scale structure (a description of this type of validity evidence is included later in this section) as well as the extent to which different scales on the instrument have been demonstrated to make unique (i.e., non-overlapping) contributions to the prediction of relevant criterion measures. These assessments were based on similar criteria to those that are described in this appendix for assessing the psychometric properties of the individual scales that were selected for review on each instrument. The resulting overall assessments of reliability and validity evidence for each instrument that are reported in this guide were made using the same 9-point scale that was used in making the parallel assessments for individual scales, as described in this Appendix. An assessment of "Limited," for example, would correspond to a rating of 3, and an assessment of "Moderate to substantial" would correspond to a rating of 6. The process used in arriving at the ratings of reliability and validity evidence for instruments, however, was less systematic and structured than that used for individual scales. Accordingly, the assessments that are provided should be regarded as having the potential to be broadly informative only.

<sup>2</sup> In developing our framework and approach, we found it helpful to consult prior efforts to evaluate the psychometric properties of measures. These resources included the *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions* prepared by Mathematica Policy Research. (see in particular [Volume II: Technical Details, Measure Profiles, and Glossary \(Appendices A – G\)](#), Malone et al., 2010) and the [Compendium of Preschool Through Elementary School Social-Emotional Learning and Associated Assessment Measures](#) prepared by the Social and Emotional Learning Group of the Coalition for Academic, Social, Emotional Learning (CASEL) at the University of Illinois at Chicago (Denham, Ji, & Hamre, 2010).

**Figure 1: Overview of Framework for Ratings of Reliability and Validity Evidence**

		Evidence Ratings					
Facet of Reliability or Validity	Orienting Questions	Quantity/ Amount <sup>a</sup>	Quality/ Rigor <sup>b</sup>	Breadth/ Comprehensiveness <sup>c</sup>	Strength <sup>d</sup>	Consistency <sup>e</sup>	Overall
<b>Reliability</b>							
Internal Consistency	Relevant?						
Inter-rater	Relevant?						
Test-retest	Relevant?						
<b>Overall</b>							
<b>Validity</b>							
Convergent	Equivalent/ highly similar constructs?						
Discriminant	Distinct constructs?						
Criterion-Related	Relevant criterion or outcome measures?						
Construct	Relevant theoretical predictions /models?						
<b>Overall</b>							

<sup>a</sup> **1**=None; **2**=Limited; **3**=Moderate; **4**=Substantial; **5**=Extensive.

<sup>b</sup> **1**=Poor; **2**=Fair; **3**=Good; **4**=Very Good; **5**=Excellent.

<sup>c</sup> **1**=None; **2**=Limited; **3**=Some; **4**=Most; **5**=All or Nearly All.

<sup>d</sup> **1**=Very Low; **2**=Low; **3**=Moderate; **4**=High; **5**=Very High.

<sup>e</sup> **1**=Highly Inconsistent; **2**=Moderately Inconsistent; **3**=[No Label]; **4**=Moderately Consistent; **5**=Highly Consistent.

Next, having answered these questions, we evaluated the available evidence as it pertained to each of several different facets of reliability and validity (see Figure 1). In the primary report, *From Soft Skills to Hard Data*, a brief explanation of each of these types of reliability and validity is provided in the section titled *Psychometrics: What Are They and Why Are They Useful?* Orienting questions similar to those listed above were used to facilitate ratings of the available evidence as it related to each aspect of a scale's reliability and validity. In the case of reliability, these questions were used to identify which aspects of reliability were relevant for a particular scale. For example, if a scale was intended to be completed only as a self-report measure by youth themselves, interrater reliability was not a relevant consideration. In the case of validity, the orienting questions focused on the specific types of evidence that would be most relevant in evaluating a particular scale's validity. For criterion-related validity, for example, we made an effort to identify the kinds of youth outcomes most likely to be influenced by the skill or concept that a scale was intended to measure. For each facet of reliability (as applicable) and validity, we evaluated the available evidence along each of several dimensions. These dimensions included:

- quantity or amount (for example, the number of different studies)
- quality and rigor (for example, when assessing convergent validity evidence, the extent to which the other scales involved had well-established validity for measuring the same skill or attribute)
- breadth and comprehensiveness (the extent to which evidence was available for particular groups such as male and female youth and, as applicable, different raters such as teachers and OST program staff)
- strength (the level of support that findings typically provided for whatever facet of reliability or validity was being considered)
- consistency (the degree to which findings were consistent across different studies or research samples).

The evidence as it related to each of these dimensions for a given facet of reliability or validity for a scale was assigned a rating from 1 to 5. (The anchor terms used for each set of ratings are noted in Figure 1.) Guidelines were developed to facilitate the assignment of these ratings for different facets of reliability and validity. Illustratively, for rating the strength of evidence for internal consistency reliability, guidelines focused on Cronbach alpha coefficient (Very Low: < .30; Low: .30-.50; Moderate: .50-.70; High: .70-.90; Very High: >.90). It should be noted, however, that in most instances guidelines were more qualitative in nature and thus required more subjective judgment in their application. Illustratively, in assessing the quality and rigor of evidence for criterion-related validity, we took into account the number and range of criterion or outcome measures, the extent to which the criterion measures were well validated, whether the measures assessed outcomes that were plausible and of likely interest for the scale, whether outcomes were assessed concurrently or at a later point in time, whether analyses included statistical control for extraneous influences, and how representative the samples involved were of the population of youth for which use of the scale was intended.

Having made ratings for each of the above dimensions for a given facet of a scale's reliability or validity, an overall rating of the evidence was assigned on a scale ranging from 1 to 9 (1 = Not at All; 3 = Limited; 5 = Moderate; 7 = Substantial; 9 = Extensive). By virtue of the different dimensions that we used to evaluate the available evidence, these ratings tended to be a function of both the scope and quality of the available evidence and the extent to which the findings obtained were supportive of the relevant facet of reliability or validity. More specifically, whereas a high rating typically required both a relative abundance of evidence and supportive findings, a low rating could be assigned either because of a general absence of evidence or because evidence was available, but it was not supportive.

The final step in the process was then to assign overall ratings of the evidence to support the scale's reliability and validity, respectively, using the same nine-point scale. These ratings served as the basis for the assessments of each scale's reliability and validity evidence that are included in this guide. An assessment of "Limited," for example, would correspond to a rating of 3, and an assessment of "Moderate to substantial" would correspond to a rating of 6.

Several considerations should be kept in mind with regard to our overall ratings of reliability and validity evidence for scales. First, these summative ratings were not arrived at by a simple averaging of the ratings provided for different facets of reliability or validity. Rather, there was room for subjective judgment to play a role, based on the totality of the available evidence. Illustratively, if ratings for a scale were at least moderately favorable across all facets of validity, this allowed us to take into account the consistency and breadth of the available evidence as an additional strength in arriving at a summative or overall rating of validity. Second, we tended to give greater weight to those facets of reliability and validity for which sufficient evidence was available to make a reasonably informed assessment. So, for example, if a scale's internal consistency reliability had been investigated extensively, but no studies had examined its test-retest reliability, our overall assessment of reliability tended to be influenced more by our rating of the former facet of reliability than the latter. In a general sense, this approach reflected our view that it was appropriate to give more weight to data that were present than data that were missing and unknown. Finally, as we have noted was the case for our ratings of specific facets of reliability and validity, our overall ratings of evidence in each area were nonetheless inevitably influenced by both the scope/quality and supportiveness of the available evidence. For this reason, assessments of reliability and validity evidence for scales reviewed in this guide that fall at the lower end of the rating scale should be interpreted with particular caution and not be taken necessarily as an indication of a scale's lack of promise or potential. In these instances, users are encouraged to take special care to also review the technical summaries that are provided for each scale so as to have an appropriate context for the summative ratings.

All ratings were arrived at independently by two of the authors of this guide (DuBois and Ji) with discrepancies resolved by conference. For the most part, there was fairly strong agreement in the ratings, especially with respect to the overall assessments of reliability and validity evidence that are reported in this guide. However, a formal assessment of interrater reliability was not conducted. Furthermore, the validity of the rating system itself has not been evaluated.<sup>1</sup> In keeping with the theme of this guide, we would thus encourage users to regard the assessments that we provide as tentative and by no means definitive or firmly established.

<sup>1</sup> For the second edition of the guide, Barb Hillaker arrived at the new ratings with review by Dubois.

# Technical Summaries of Scales Reviewed from the Resilience & Youth Development Module of the California Healthy Kids Survey (CHKS)

## Empathy

The Empathy scale of the Middle and High School versions of the *CHKS Resilience & Youth Development Module (RYDM)* is a three-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. To date, evidence for the scale's reliability is moderate to substantial, and evidence for the scale's validity is moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within several large samples of 7th-, 9th- and 11th-grade students in California school districts. Internal consistency reliability has been satisfactory in these samples for youth at each grade level, male and female youth, and White, African-American, Mexican-American, and Chinese-American youth and was also found to be satisfactory for a sample of 1,257 students attending alternative high schools in California. In a sample of 90 ninth-grade students from seven classrooms in two schools in a large urban school district, a moderate level of test-retest reliability was found over a two-week interval ( $r = .57$ ).

### Validity

In support of the scale's discriminant validity, in exploratory factor analyses of personal resilience items from the RYDM with a sample of 12,000 grade 7, 9 and 11 students in California school districts (as well as an equal size validation sample), the items from the scale loaded together on the same factor distinct from those intended to represent other constructs (Cooperation and Communication, Goals and Aspirations, Self-Efficacy, Self-Awareness, and Problem-Solving. Similar findings were obtained in factor analyses of both personal and environmental resilience items from the RYDM with large samples of California school students in grades 7, 9 and 11.) Confirmatory factor analysis replicated this finding, although the factor representing the scale in this analysis demonstrated substantial overlap with other factors intended to represent Self-Efficacy, Problem-Solving, and Self-Awareness ( $r_s = .69$  to  $.82$ ). Scores for the latent factor representing the scale, in turn, exhibited significant associations in expected directions with youth responses to items on the CHKS core survey asking about substance use ( $r_s = -.11$  to  $-.23$ ), violence (significant  $r_s = -.05$  to  $-.26$ ), grades ( $r = .22$ ), and truancy ( $r = -.15$ ). For a sample of 651 students in a large county in Southern California, the same factor score exhibited a weak significant association with higher academic achievement test scores in English/language arts ( $r = .09$ ), but a non-significant association with scores in the area of math.

### Summary

This scale shows evidence of reliability across demographically varied groups of youth. Factor analysis findings offer support for the scale as a measure of empathy distinct from other constructs. To add to these results, it would be useful to investigate the scale's associations with well-validated measures of empathy (convergent validity) as well as its relative degree of overlap with measures of skills or abilities in other areas (discriminant validity). The scale's concurrent associations with youth self-reported behavioral and academic outcomes provide encouraging evidence of criterion validity. It would be helpful in future research both to expand examination of the scale's associations with outcome measures derived from other sources and to consider outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs also would be valuable. Furthermore, although developed for use with secondary school students, it could be worthwhile to explore the scale's viability as a measure for younger, elementary school-age students. (The Elementary School version of the CHKS already includes a scale comprising two of the three items from the scale.)

## Cooperation and Communication

The Cooperation and Communication scale of the *CHKS* Middle and High School versions of the *Resilience & Youth Development Module (RYDM)* is a three-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for reliability is moderate. As discussed below, currently available evidence is not supportive of the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within large samples of 7th-, 9th- and 11th-grade students in California school districts, as well as a sample of 1,257 students attending alternative high schools in California. Data on test-retest reliability are not available.

### Validity

In exploratory factor analyses of personal resilience items from the *RYDM* with a sample of 12,000 grade 7, 9 and 11 students in California school districts (as well as an equal size validation sample), the items comprising the scale failed to load together on the same factor (one item loaded on a factor with items intended to assess Self-Efficacy and Problem-Solving and the other two items each had loadings on multiple factors). In further exploratory factor analyses of both personal and environmental resilience items from the *RYDM* with the samples of grade 7, 9 and 11 students referred to above under Reliability, the scale items similarly loaded on the same factor as those from Self-Efficacy scale.

### Summary

Reliability evidence for this scale is encouraging. Available validity evidence, however, indicates a potential need for revisions to scale content to more specifically assess the targeted construct of cooperation and communication as distinct from other constructs. To help clarify this issue, it would be useful in future research to investigate the scale's association with other well-validated measures of communication and/or cooperation (convergent validity) as well as its relative degree of overlap with measures that target skills or abilities in other areas (discriminant validity).



## Problem-Solving

The Problem-Solving scale of the Middle and High School versions of the *CHKS Resilience & Youth Development Module (RYDM)* is a three-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Critical Thinking and Decision-Making. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within large samples of 7th-, 9th- and 11th-grade students in California school districts, as well as a sample of 1,257 students attending alternative high schools in California. Data on test-retest reliability are not available.

### Validity

In support of discriminant validity, in exploratory factor analyses of personal and environmental resilience items from the *RYDM* with large samples of students in grades 7, 9 and 11 in California school districts, the items from the scale loaded together on the same factor distinct from those intended to represent other constructs. In further exploratory factor analyses of only the personal resilience items on the *RYDM* with a sample of 12,000 grade 7, 9 and 11 students in California school districts (as well as an equal size validation sample), the items from the scale for the most part similarly loaded together on the same factor; however, items intended to assess other constructs (most notably, Self-Efficacy) also loaded highly on the same factor. In confirmatory factor analyses, the best-fitting model included a factor comprised of two of the three items from the scale. This factor demonstrated substantial overlap with other factors intended to represent Empathy, Self-Efficacy and Self-Awareness ( $r_s = .62$  to  $.82$ ). Scores on the corresponding latent factor, in turn, exhibited significant associations in expected directions with youth responses to items on the *CHKS* core survey asking about substance use ( $r_s = -.17$  to  $-.29$ ), violence ( $r_s = -.08$  to  $-.20$ ), depressed mood ( $r = -.11$ ), grades ( $r = .21$ ) and truancy ( $r = -.17$ ). For a sample of 651 students in a large county in Southern California, however, the same factor score did not exhibit significant associations with academic achievement test scores in English/language arts or math.

### Summary

Reliability evidence for this scale is encouraging. Factor analysis findings, although generally encouraging, are not entirely consistent in offering support for the scale as a measure of problem-solving distinct from other resilience constructs. In future research, it would be useful to investigate the scale's associations with well-validated measures of problem-solving ability (convergent validity) as well as its relative degree of overlap with measures of skills or abilities in other areas (discriminant validity). The concurrent associations of a factor score based on all but one of the items from the scale with youth self-reported behavioral and academic outcomes provide encouraging evidence of criterion validity. To add to findings in this area, it would be helpful both to expand examination of the scale's associations with outcome measures derived from other sources and to consider outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs also would be valuable. Furthermore, although developed for use with secondary school students, it would be worthwhile to explore the scale's viability as a measure for younger, elementary school-age students. (The Elementary School version of the *CHKS* already includes a scale comprising two of the three items from the scale.)

## Self-Awareness

The Self-Awareness scale of the Middle and High School versions of the *CHKS Resilience & Youth Development Module (RYDM)* is a three-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within several large samples of 7th-, 9th- and 11th-grade students in California school districts. Internal consistency reliability has been satisfactory in these samples for youth at each grade level, male and female youth, and White, African-American, Mexican-American, and Chinese-American youth, and was also found to be satisfactory for a sample of 1,257 students attending alternative high schools in California. In a sample of 90 ninth-grade students from seven classrooms in two schools in a large urban school district, a substantial level of test-retest reliability was found over a two-week interval ( $r = .71$ ).

### Validity

In support of the scale's discriminant validity, in exploratory factor analyses of personal resilience items from the *RYDM* with a sample of 12,000 grade 7, 9 and 11 students in California school districts (as well as an equal size validation sample), the items from the scale loaded together on the same factor, distinct from those intended to represent other constructs (Cooperation and Communication, Goals and Aspirations, Self-Efficacy, Empathy and Problem-Solving. Similar findings were obtained in factor analyses of both personal and environmental resilience items from the *RYDM* with large samples of California school students in grades 7, 9 and 11.) Confirmatory factor analysis replicated this finding, although the factor representing the scale in this analysis demonstrated substantial overlap with other factors intended to represent Empathy, Problem-Solving and Self-Efficacy ( $r_s = .62$  to  $.82$ ). Scores for the latent factor representing the scale, in turn, exhibited significant associations in expected directions with youth responses to items on the *CHKS* core survey asking about substance use ( $r_s = -.14$  to  $-.25$ ), violence ( $r_s = -.10$  to  $-.24$ ), depressed mood ( $r = -.30$ ) and truancy ( $r = -.19$ ). Higher scores were correlated significantly with lower self-reported grades in school ( $r = -.20$ ), which is not necessarily expected. Furthermore, for a sample of 651 students in a large county in Southern California, the same factor score did not exhibit significant associations with academic achievement test scores in English/language arts or math.

### Summary

This scale shows encouraging evidence of reliability across demographically varied groups of youth. Factor analysis findings offer support for the scale as a measure of self-awareness distinct from other resilience constructs. To add to these results, it would be useful to investigate the scale's associations with well-validated measures of self-awareness (convergent validity), as well as its relative degree of overlap with measures of skills or abilities in other areas (discriminant validity). The scale's concurrent associations with youth self-reported behavioral and academic outcomes provide encouraging evidence of criterion validity. It would be helpful in future research both to expand examination of the scale's associations with outcome measures derived from other sources and to consider outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs also would be valuable. Furthermore, although developed for use with secondary school students, it could be worthwhile to explore the scale's viability as a measure for younger, elementary school-age students.

## Self-Efficacy

The Self-Efficacy scale of the Middle and High School versions of the *CHKS Resilience & Youth Development Module (RYDM)* is a three-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within large samples of 7th-, 9th- and 11th-grade students in California school districts, as well as a sample of 1,257 students attending alternative high schools in California. Data on test-retest reliability are not available.

### Validity

In exploratory factor analyses of personal resilience items from the *RYDM* with a sample of 12,000 grade 7, 9 and 11 California students (as well as an equal size validation sample), the items from the scale for the most part loaded together on the same factor. However, items intended to assess Problem-Solving on the *RYDM* and the item intended to assess Communication and Cooperation referred to above also loaded highly on the same factor. (In factor analyses of both personal and environmental resilience items from the *RYDM* with the samples of students in grades 7, 9 and 11 referred to above under Reliability, the scale items similarly loaded on the same factor as those from the Communication and Cooperation scale.) In confirmatory factor analyses, the best-fitting model included a factor composed of the three items from the scale, but also the Communication and Cooperation item. This factor demonstrated substantial overlap with other factors intended to represent Empathy, Problem-Solving and Self-Awareness ( $r_s = .73$  to  $.82$ ). Scores on the corresponding latent factor, in turn, exhibited significant associations in expected directions with youth responses to items on the CHKS core survey asking about substance use ( $r_s = -.17$  to  $-.29$ ), violence ( $r_s = -.11$  to  $-.25$ ), depressed mood ( $r = -.30$ ), grades ( $r = .29$ ) and truancy ( $r = -.20$ ). For a sample of 651 students in a large county in Southern California, however, the same factor score did not exhibit significant associations with academic achievement test scores in English/language arts or math.

### Summary

Reliability evidence for this scale is encouraging. Factor analysis findings offer only partial support for the scale as a measure of problem-solving distinct from other resilience constructs. In future research, it would be useful to investigate the scale's associations with well-validated measures of self-efficacy beliefs (convergent validity), as well as its relative degree of overlap with measures of other types of youth attributes (discriminant validity). The concurrent associations of a factor score based primarily on items from the scale with youth self-reported behavioral and academic outcomes provide encouraging evidence of criterion validity. To add to findings in this area, it would be helpful both to expand examination of the scale's associations with outcome measures derived from other sources and to consider outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs also would be valuable. Furthermore, although developed for use with secondary school students, it would be worthwhile to explore the scale's viability as a measure for younger, elementary school-age students.

# Technical Summaries of Scales Reviewed from the *Developmental Assets Profile (DAP)*

## Commitment to Learning

The Commitment to Learning scale of the *Developmental Assets Profile* is a seven-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is substantial and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within multiple samples, including a sample of approximately 1,300 sixth- through 12th-grade students from a public school district in Minnesota and a sample of approximately 1,110 sixth- through 8th-grade students attending two public middle schools in Oregon. Findings for these samples include satisfactory internal reliability within subgroups of youth defined by gender, age (grade 6 through 8 vs. grade 9 through 12) and race/ethnicity (White, Hispanic, Asian-American, American Indian and Multi-racial). For the Minnesota sample, test-retest reliability assessed over a two-week interval for a subsample of 200 youth also was high both for the group as a whole ( $r = .84$ ) and for gender and age subgroups ( $r_s > .80$ ).

### Validity

In support of convergent validity, for the Minnesota sample referred to above, scores on the scale were correlated positively with the number of assets in the same domain as assessed on the Search Institute's *Profiles of Student Life: Attitudes and Behavior* survey ( $r_s > .60$  for overall sample, as well as for boys and girls and students in grades 6 through 8 and 9 through 12). In the Oregon sample referred to above, the two middle schools that youth attended were in the same district but differed substantially in the quality and quantity of resources available to youth based on the views of school personnel, youth workers, community leaders and parents. As would be predicted, youth attending the more "asset-rich" school had significantly higher scores on the scale than those attending the less "asset-rich" school. In an evaluation of an intervention intended to develop the social and financial competencies of adolescent girls in Bangladesh involving 498 girls ages 10-18 whose villages had been randomly assigned to receive the intervention or not, those who participated in the intervention demonstrated evidence of greater positive change in their scores on the scale over a six-month period relative to those in the control group.

### Summary

Evidence supports the scale's reliability for demographically varied groups of youth. With regard to validity, available findings do not clearly establish the scale as a measure of commitment to learning distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of investment in learning (convergent validity) relative to measures that target attitudes or skills in other areas (discriminant validity). Because several of the items on the scale reference school, it would be useful in particular to clarify support for the scale as a measure of interest in learning generally as contrasted with academic motivation more specifically. The scale has demonstrated evidence of sensitivity to effects of OST program participation in an international context. It would be useful in future research to add to findings in this area by examining sensitivity to program effects within U.S. samples. It also would be helpful to have evidence regarding the scale's ability to predict outcomes of interest such as academic achievement (criterion validity).

## Social Competencies

The Social Competencies scale of the *Developmental Assets Profile* is an eight-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is substantial and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within multiple samples, including a sample of approximately 1,300 sixth- through 12th-grade students from a public school district in Minnesota and a sample of approximately 1,110 sixth- through 8th-grade students attending two public middle schools in Oregon. Findings for these samples include satisfactory reliability within subgroups of youth defined by gender, age (grade 6 through 8 vs. grade 9 through 12) and race/ethnicity (White, Hispanic, Asian-American, American Indian and Multi-racial). For the Minnesota sample, test-retest reliability assessed over a two-week interval for a subsample of 200 youth also was high both for the group as a whole ( $r = .81$ ) and for gender and age subgroups ( $r_s > .80$ ).

### Validity

In support of convergent validity, for the Minnesota sample referred to above, scores on the scale were correlated positively with the number of assets in the same domain as assessed on the Search Institute's *Profiles of Student Life: Attitudes and Behavior* survey ( $r_s > .60$  for overall sample, as well as boys and girls and students in grades 6 through 8 and 9 through 12). In the Oregon sample, the two middle schools that youth attended were in the same school district but differed substantially in the quality and quantity of resources available based on the views of school personnel, youth workers, community leaders and parents. As would be predicted, youth attending the more "asset-rich" school had significantly higher scores on the scale than those attending the less "asset-rich" school. In an evaluation of an intervention intended to develop the social and financial competencies of adolescent girls in Bangladesh involving 498 girls ages 10 to 18 whose villages had been randomly assigned to receive the intervention or not, those who participated in the intervention demonstrated evidence of greater positive change in their scores on the scale over a six-month period relative to those in the control group.

### Summary

Evidence supports the scale's reliability for demographically varied groups of youth. With regard to validity, available findings do not clearly establish the scale as a measure of social competence distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of abilities in this domain (convergent validity) relative to those measures that target skills in other areas such as academics (discriminant validity). The scale has demonstrated evidence of sensitivity to effects of OST program participation in an international context. It would be useful in future research to add to findings in this area by examining sensitivity to program effects within U.S. samples. It also would be helpful to have evidence regarding the scale's ability to predict outcomes of interest such as those relating to mental health or future employment success (criterion validity).

## Positive Identity

The Positive Identity scale of the *Developmental Assets Profile* is a six-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within multiple samples, including a sample of approximately 1,300 sixth- through 12th-grade students from a public school district in Minnesota and a sample of approximately 1,110 sixth- through 8th-grade students attending two public middle schools in Oregon. Findings for these samples include satisfactory reliability within subgroups of youth defined by gender, age (grade 6 through 8 vs. grade 9 through 12) and race/ethnicity (White, Hispanic, Asian-American, American Indian, and Multi-racial). For the Minnesota sample, test-retest reliability assessed over a two-week interval for a subsample of 200 youth also was satisfactory both for the group as a whole ( $r = .78$ ) and for gender and age subgroups ( $r_s > .70$ ).

### Validity

In support of convergent validity, for the Minnesota sample referred to above, scores on the scale were correlated positively with the number of assets in the same domain as assessed on the Search Institute's *Profiles of Student Life: Attitudes and Behavior* survey ( $r_s > .60$  for overall sample, as well as boys and girls and students in grades 6 through 8 and 9 through 12). In further support of the scale's convergent validity, in the Oregon sample the scale exhibited strong correlations with two established measures of self-esteem, the Global Self-Worth Scale of the *Self-Perception Profile for Adolescents* ( $r = .72$ ) and the *Rosenberg Self-Esteem Scale* ( $r = .72$ ). The two middle schools that youth in this sample attended were in the same school district but differed substantially in the quality and quantity of resources available, based on the views of school personnel, youth workers, community leaders and parents. As would be predicted, youth attending the more "asset-rich" school had significantly higher scores on the Positive Identity scale than those attending the less "asset-rich" school. In an evaluation of an intervention intended to develop the social and financial competencies of adolescent girls in Bangladesh involving 498 girls ages 10-18 whose villages had been randomly assigned to receive the intervention or not, those who participated in the intervention demonstrated evidence of greater positive change in their scores on the scale over a six-month period relative to those in the control group.

### Summary

Evidence supports the scale's reliability for demographically varied groups of youth. Validity evidence includes expected associations with measures of a related construct (self-esteem). Available findings, however, do not clearly establish the scale as a measure of positive identity distinct from attitudes or skills in other areas (discriminant validity). A majority of the items on the scale refer to effective coping with challenges or sustaining a positive orientation toward the future. This suggests the usefulness of examining support for the scale as a measure of positive identity distinct from other attributes such as self-efficacy or optimism. The scale has demonstrated evidence of sensitivity to effects of OST program participation in an international context. It would be useful in future research to add to findings in this area by examining sensitivity to program effects within U.S. samples. It also would be helpful to have evidence regarding the scale's ability to predict youth outcomes of interest (criterion validity).

# Technical Summaries of Scales Reviewed from the *Devereux Student Strengths Assessment (DESSA)*

## Self-Awareness

The Self-Awareness scale of the *DESSA* is a seven-item measure designed to be completed by parents, teachers, or staff in settings such as OST programs. For purposes of this guide, consideration is limited to findings obtained using teachers and staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was satisfactory ( $r = .86$ ), as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide) of students at unspecified grade levels ( $r = .72$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, scores on the scale exhibited moderate to strong correlations with the Intrapersonal Strengths ( $r = .59$ ), Affective Strengths ( $r = .59$ ) and Interpersonal Strengths ( $r = .49$ ) scales of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Adaptability ( $r = .67$ ), Social Skills ( $r = .72$ ), Leadership ( $r = .76$ ) and Study Skills ( $r = .69$ ) scales of the *Behavior Assessment System for Children-2 (BASC-2)*. As these scales target constructs that are largely conceptually distinct from self-awareness, the substantial magnitude of the correlations does not suggest a high level of discriminant validity for the scale. Furthermore, the scale's association with the most closely aligned of these other scales (Intrapersonal Strengths) is not consistently stronger than those that it exhibited with the remaining scales. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning ( $r = .56$ ) scale of the *BERS-2* and the Aggression ( $r = -.69$ ), Conduct Problems ( $r = -.70$ ) and Depression ( $r = -.59$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is limited to a single study. Available findings do not clearly establish the scale as a measure of self-awareness distinct from other constructs. To address this concern, in future research it would be useful to investigate the scale's associations with well-validated measures of self-awareness relative to measures of youth attributes or skills in other areas. The scale's concurrent associations with teacher-report indices of youth academic, behavioral and emotional functioning provide encouraging evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.

## Social Awareness

The Social Awareness scale of the *DESSA* is a nine-item measure completed by parents, teachers or staff in settings such as OST programs. For purposes of this guide, consideration is limited to findings obtained using teachers or staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was satisfactory ( $r = .93$ ) as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide), also of students at unspecified grade levels ( $r = .70$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, scores on the scale exhibited moderate to strong correlations with the Interpersonal Strengths ( $r = .74$ ), Affective Strength ( $r = .49$ ) and Intrapersonal Strengths ( $r = .53$ ) scales of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Adaptability ( $r = .84$ ), Social Skills ( $r = .78$ ), Leadership ( $r = .76$ ) and Study Skills ( $r = .66$ ) scales of the *Behavior Assessment System for Children-2 (BASC-2)*. The pattern of these correlations tends to involve relatively stronger correlations of the scale with other constructs that involve social components (Interpersonal Strengths, Social Skills, Leadership) and as such provides some support for the scale's convergent and discriminant validity. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning scale of the *BERS-2* ( $r = .56$ ) and the Aggression ( $r = -.69$ ), Conduct Problems ( $r = -.70$ ) and Depression ( $r = -.59$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal, and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is also encouraging but is limited to a single study. Further data are needed to clarify the degree to which the scale specifically assesses social awareness as distinct from other constructs, especially those that also include a social component (e.g., leadership). To address this concern, it would be useful to investigate the scale's strength of association with well-validated measures of social awareness relative to measures of skills or abilities in other areas (convergent and discriminant validity). The scale's concurrent associations with teacher-report indices of youth academic, behavioral and emotional functioning provide evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.



## Self-Management

The Self-Management scale of the *DESSA* is an 11-item measure designed to be completed by parents, teachers, or staff in settings such as after-school programs. For purposes of this guide, consideration is limited to findings obtained using teachers and staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is moderate as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was satisfactory ( $r = .93$ ), as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide) of students at unspecified grade levels ( $r = .75$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, Interpersonal Strengths ( $r = .75$ ), Intrapersonal Strengths ( $r = .61$ ) and Affective Strengths ( $r = .50$ ) scales of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Adaptability ( $r = .87$ ), Social Skills ( $r = .77$ ), Leadership ( $r = .76$ ) and Study Skills ( $r = .72$ ) scales of the *Behavior Assessment System for Children-2*. The scale's relatively stronger associations with the scales on these other measures that are most closely aligned with self-management skills (Interpersonal Strengths, Adaptability) provides some evidence for the scale's convergent and discriminant validity. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning scale of the *BERS-2* ( $r = .69$ ) and the Aggression ( $r = -.63$ ), Conduct Problems ( $r = -.69$ ), Anxiety ( $r = -.30$ ) and Depression ( $r = -.51$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal, and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is also encouraging but is limited to a single study. Further data are needed to clarify the degree to which the scale specifically assesses self-management as distinct from other constructs. To address this concern, it would be useful to investigate the scale's strength of association with well-validated measures of self-management relative to measures of skills or abilities in other areas (convergent and discriminant validity). The scale's concurrent associations with teacher-report indices of youth academic, behavioral and emotional functioning provide evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.

## Goal-Directed Behavior

The Goal-Directed Behavior scale of the *DESSA* is a 10-item measure completed by parents, teachers, or staff in settings such as after-school programs. For purposes of this guide, consideration is limited to findings obtained using teachers or staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was satisfactory ( $r = .90$ ), as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide) of students at unspecified grade levels ( $r = .77$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, scores on the scale exhibited moderate to strong correlations with the Interpersonal Strengths ( $r = .67$ ) and Intrapersonal Strengths ( $r = .67$ ) scales of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Adaptability ( $r = .70$ ), Social Skills ( $r = .71$ ), Leadership ( $r = .80$ ) and Study Skills ( $r = .82$ ) scales of the *Behavior Assessment System for Children-2 (BASC-2)*. As these scales assess constructs that are conceptually distinct from goal-directed behavior, the substantial magnitude of the correlations does not suggest a high level of discriminant validity for the scale. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning ( $r = .72$ ) and Affective Strength ( $r = .59$ ) scales of the *BERS-2* and the Aggression ( $r = -.47$ ), Conduct Problems ( $r = -.57$ ) and Depression ( $r = -.42$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is limited to a single study. Available findings do not clearly establish the scale as a measure of goal-directed behavior distinct from other constructs. To address this concern, in future research it would be useful to investigate the scale's associations with well-validated measures of goal-directed behavior relative to measures of skills or behaviors in other areas. The scale's concurrent associations with teacher-report indices of youth academic, behavioral, and emotional functioning provide encouraging evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.

## Relationship Skills

The Relationship Skills scale of the *DESSA* is a 10-item measure designed to be completed by parents, teachers, or staff in settings such as after-school programs. For purposes of this guide, consideration is limited to findings obtained using teachers and staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is moderate as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was satisfactory ( $r = .92$ ), as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide) of students at unspecified grade levels ( $r = .71$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, scores on the scale exhibited moderate to strong correlations with the Interpersonal Strengths ( $r = .72$ ) scale of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Social Skills ( $r = .85$ ) scale of the Behavioral Assessment System for Children (*BASC-2*). The former association was somewhat stronger than the scale's correlation with the Intrapersonal Strengths and Affective Strengths scales of the *BERS-2* ( $r_s = .63$ ) and the latter association similarly was somewhat stronger than the scale's correlation with the Adaptability ( $r = .81$ ), Leadership ( $r = .77$ ) and Study Skills ( $r = .66$ ) scales of the *BASC-2*. This pattern of relatively stronger associations with other measures of social competence (Interpersonal Strengths, Social Skills) provides some evidence for the scale's convergent and discriminant validity. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning ( $r = .55$ ) and Affective Strength ( $r = .63$ ) scales of the *BERS-2* and the Aggression ( $r = -.62$ ), Conduct Problems ( $r = -.66$ ) and Depression ( $r = -.51$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is also encouraging but is limited to a single study. Further data are needed to clarify the degree to which the scale specifically assesses relationship skills as distinct from other constructs. To address this concern, it would be useful to investigate the scale's strength of association with well-validated measures of social competence relative to measures of skills or abilities in other areas (convergent and discriminant validity). The scale's concurrent associations with teacher-report indices of youth academic, behavioral, and emotional functioning provide evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.

## Personal Responsibility

The Personal Responsibility scale of the *DESSA* is a 10-item measure designed to be completed by parents, teachers or staff in settings such as OST programs. For purposes of this guide, consideration is limited to findings obtained using teachers or staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was high ( $r = .94$ ), as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide) of students at unspecified grade levels ( $r = .92$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, scores on the scale exhibited moderate to strong correlations with the Interpersonal Strengths ( $r = .66$ ), Affective Strength ( $r = .49$ ) and Intrapersonal Strengths ( $r = .63$ ) scales of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Adaptability ( $r = .71$ ), Social Skills ( $r = .75$ ), Leadership ( $r = .84$ ) and Study Skills ( $r = .81$ ) scales of the *Behavior Assessment System for Children-2 (BASC-2)*. As these scales assess constructs that are conceptually distinct from personal responsibility, the substantial magnitude of the correlations does not suggest a high level of discriminant validity for the scale. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning scale of the *BERS-2* ( $r = .78$ ) and the Aggression ( $r = -.50$ ), Conduct Problems ( $r = -.59$ ) and Depression ( $r = -.43$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is limited to a single study. Available findings do not clearly establish the scale as a measure of personal responsibility distinct from other constructs. To address this concern, in future research it would be useful to investigate the scale's associations with well-validated measures of personal responsibility relative to measures of skills or behaviors in other areas. The scale's concurrent associations with teacher-report indices of youth academic, behavioral, and emotional functioning provide encouraging evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.

## Decision Making

The Decision-Making scale of the *DESSA* is an eight-item measure completed by parents, teachers or staff in settings such as after-school programs. For purposes of this compendium, consideration is limited to findings obtained using teachers and staff as informants. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Critical Thinking and Decision-Making. Evidence for the scale's reliability is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a national sample of ratings from 781 teachers of students in grades K-8. Test-retest reliability over a period of four to eight days for ratings provided by a sample of 38 teachers for students at unspecified grade levels was high ( $r = .94$ ), as was interrater reliability for a sample of 51 pairs of teachers (or a teacher and teacher aide) of students at unspecified grade levels ( $r = .84$ ). Reliability estimates for ratings of different demographic subgroups of youth are not available.

### Validity

In a study involving 94 teachers who provided ratings of students at unspecified grade levels, scores on the scale exhibited moderate to strong correlations with the Interpersonal Strengths ( $r = .75$ ), Affective Strength ( $r = .58$ ) and Intrapersonal Strengths ( $r = .63$ ) scales of the *Behavioral and Emotional Rating Scales-2 (BERS-2)* and the Adaptability ( $r = .79$ ), Social Skills ( $r = .79$ ), Leadership ( $r = .81$ ) and Study Skills ( $r = .74$ ) scales of the *Behavior Assessment System for Children-2 (BASC-2)*. As these scales assess constructs that are conceptually distinct from decision-making, the substantial magnitude of the correlations does not suggest a high level of discriminant validity for the scale. In support of criterion validity, in the same study scores on the scale exhibited significant associations in expected directions with the School Functioning scale of the *BERS-2* ( $r = .69$ ) and the Aggression ( $r = -.66$ ), Conduct Problems ( $r = -.73$ ) and Depression ( $r = -.52$ ) scales of the *BASC-2*. As also would be expected, youth with higher ratings on the Family Involvement scale of the *BERS-2* tended to be rated more favorably on the scale, and those with ratings indicative of more difficulties on Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales on the *BASC-2* tended to receive lower ratings.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). Validity evidence is limited to a single study. Available findings do not clearly establish the scale as a measure of decision-making distinct from other constructs. To address this concern, in future research it would be useful to investigate the scale's associations with well-validated measures of decision-making relative to measures of skills or behaviors in other areas. The scale's concurrent associations with teacher-report indices of youth academic, behavioral, and emotional functioning provide encouraging evidence of criterion validity. It would be helpful to build on these findings by examining the scale's associations with outcome measures derived from other sources, as well as outcomes assessed at later points in youths' schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable, as would collecting data on reliability and validity when ratings are provided by program staff rather than teachers.

# Technical Summaries of Scales Reviewed from the *Holistic Student Assessment (HSA)*

## Assertiveness

The Assertiveness scale of the *Holistic Student Assessment (HSA)* is a 13-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate, and evidence for the scale's validity is limited.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

With relevance to the scale's discriminant validity, in an exploratory bi-factor analysis conducted with data collected from 423 youth in grades 5 through 9 attending public elementary, middle and junior high schools in a metropolitan city in the Northeast, three of the Assertiveness items loaded on the same factor as the items from the Action Orientation scale of the HSA, which is not included in this guide. Three other assertiveness items did not load significantly on any factor. With relevance to concurrent validity, for this same sample high scores on the scale exhibited an expected significant negative association with scores on the Peer Problems scale of the Strengths & Difficulties Questionnaire ( $r = -.17$ ), but not with scores on the Emotional Problems scale of the same measure ( $r = -.07$ ).

### Summary

Support for the scale's reliability is encouraging but is limited to a single sample and does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence, which is similarly limited, offers only mixed support for the scale as a measure of assertiveness. It would be useful in future research to gather information regarding the scale's association with well-validated measures of assertiveness (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to further examine the scale's expected associations with outcomes in other areas, including those assessed at later points in youths' schooling or development (predictive validity). It would be helpful as well to investigate the scale's sensitivity to effects of OST program participation.

## Emotional Control

The Emotional Control subscale of the *HSA* is a six-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate. There is no available evidence specific to the validity of this scale, which is one of seven subscales that comprise the broader Resiliencies domain on the *HSA*.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

The authors state that “attributes measured by the *HSA* assessment tool are significantly related to the SDQ [Strengths and Difficulties Questionnaire] and academic outcomes.” This would support criterion validity. However, no detailed information is available on the validity of the Emotional Control scale.

### Summary

Support for the scale's reliability is encouraging but does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence specific to this scale is not available. It would be useful in future research to gather information regarding the scale's association with well-validated measures of emotional control (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to examine the scale's expected associations with outcomes in other areas, including those assessed at later points in youths' schooling or development (predictive validity). It would be helpful as well to investigate the scale's sensitivity to effects of OST program participation.

## Perseverance

The Perseverance subscale of the *HSA* is a five-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is moderate. There is no available evidence specific to the validity of this scale, which is one of five subscales that comprise the broader Learning/School Engagement domain on the *HSA*.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 primarily economically at-risk minority youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

The authors state "attributes measured by the *HSA* assessment tool are significantly related to the SDQ [Strengths and Difficulties Questionnaire] and academic outcomes." This would support criterion validity. However, no detailed information is available on the validity of the Perseverance scale.

### Summary

Support for the scale's reliability is encouraging but does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence specific to this scale is not available. It would be useful in future research to gather information regarding the scale's association with well-validated measures of perseverance (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to examine the scale's expected associations with outcomes in other areas, including those assessed at later points in youths' schooling or development (predictive validity). It would be helpful as well to investigate the scale's sensitivity to effects of OST program participation.



## Critical Thinking

The Critical Thinking subscale of the *HSA* is a six-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Critical Thinking. Evidence for the scale's reliability is moderate. There is no available evidence specific to the validity of the scale, which is one of seven subscales that comprise the broader Resiliencies domain on the *HSA*.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 primarily economically at-risk minority youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

The authors state that “attributes measured by the *HSA* assessment tool are significantly related to the SDQ [Strengths and Difficulties Questionnaire] and academic outcomes” (Robertson, et.al, 2013), but no information is available on the validity of the Critical Thinking scale.

### Summary

Support for the scale's reliability is encouraging but does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence specific to this scale is not available. It would be useful in future research to gather information regarding the scale's association with well-validated measures of critical thinking (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to examine the scale's expected associations with outcomes in other areas, including those assessed at later points in youths' schooling or development (predictive validity). It would be helpful as well to investigate the scale's sensitivity to effects of OST program participation.

## Empathy

The Empathy scale is a six-item self-assessment measure. In the current *HSA* (version six), the previous Interpersonal Sensitivity scale has been merged with the Empathy scale. The content of this scale maps onto Relationships and Collaboration focus area for this guide. Reliability for this scale is moderate. There is no available evidence specific to the validity of this scale, which is one of seven subscales that comprise the broader Resiliencies domain on the *HSA*.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

The authors state “attributes measured by the *HSA* assessment tool are significantly related to the SDQ [Strengths and Difficulties Questionnaire] and academic outcomes” (Robertson, et.al, 2013), but no information is available on the validity of the Empathy scale.

### Summary

Support for the scale’s reliability is encouraging but does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence specific to this scale is not available. It would be useful in future research to gather information regarding the scale’s association with well-validated measures of empathy (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to examine the scale’s expected associations with outcomes in other areas, including those assessed at later points in youths’ schooling or development (predictive validity). It would be helpful as well to investigate the scale’s sensitivity to effects of OST program participation.

## Relationships with Peers

The Relationships with Peers scale of the *HSA* is a six-item self-assessment measure. The content of this scale maps onto Relationships and Collaboration focus area for this guide. Evidence for the scale's reliability is moderate. There is no available evidence specific to the validity of this scale, which is one of two subscales that comprise the broader Relationships domain on the *HSA*.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

The authors state “attributes measured by the *HSA* assessment tool are significantly related to the SDQ [Strengths and Difficulties Questionnaire] and academic outcomes” (Robertson, et.al, 2013), but no information is available on the validity of the Relationship with Peers scale.

### Summary

Support for the scale's reliability is encouraging but does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence specific to this scale is not available. It would be useful in future research to gather information regarding the scale's association with well-validated measures of relationships with peers (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to examine the scale's expected associations with outcomes in other areas, including those assessed at later points in youths' schooling or development (predictive validity). It would be helpful as well to investigate the scale's sensitivity to effects of OST program participation.

## Relationships with Adults

The Relationships with Adults scale of the *HSA* is a six-item self-assessment measure. The content of this scale maps onto Relationships and Collaboration focus area for this guide. Evidence for the scale's reliability is moderate. There is no available evidence specific to the validity of this scale, which is one of two subscales that comprise the broader Relationships domain on the *HSA*.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a sample of 1,293 youth who attended afterschool programs or were attending school in a metropolitan city in the Northeast. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest reliability.

### Validity

The authors state “attributes measured by the HSA assessment tool are significantly related to the SDQ [Strengths and Difficulties Questionnaire] and academic outcomes” (Robertson, et.al, 2013), but no information is available on the validity of the Relationship with Adults scale.

### Summary

Support for the scale's reliability is encouraging but does not include information on reliability for different demographic subgroups of youth or test-retest reliability. Validity evidence specific to this scale is not available. It would be useful in future research to gather information regarding the scale's association with well-validated measures of emotion control (convergent validity), as well as further data on its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It also would be useful to examine the scale's expected associations with outcomes in other areas, including those assessed at later points in youths' schooling or development (predictive validity). It would be helpful as well to investigate the scale's sensitivity to effects of OST program participation.

# Technical Summaries of Scales Reviewed from the *San Francisco Beacons Youth Survey*

## School Effort

The School Effort scale from the San Francisco Beacons Youth Survey (Beacons Youth Survey) is a four-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is limited and evidence for the scale's validity is moderate.

### Reliability

Within a sample of 432 youth in grades 7 and 8 from 10 Boys & Girls Clubs, internal consistency reliability of the scale was found to be moderate (coefficient alpha = .59). No data are available on test-retest reliability.

### Validity

In path modeling analyses conducted as part of a study of 838 students attending three middle schools that were implementing OST programming in the San Francisco Beacon Initiative, positive change on the scale between the fall of one school year and the spring of the following year was predictive of increased grades across the two school years. Change on the scale also mediated the associations that measures of developmental experiences (e.g., support from non-familial adults) from the fall assessment and change over time in other youth attributes (e.g., self-efficacy beliefs) exhibited with improvement in grades. In the same analyses, however, positive change on the scale was unexpectedly predictive of increased absences from school and was not found to be related to improvement on standardized test scores for reading or math. Further analyses compared the end-point spring scores on the scale for nine different groups of students who participated in varying numbers of OST program sessions (each of which lasted two to four months) that differed in their inclusion of educational and other activities with those for program non-participants, controlling for the baseline score and a range of additional variables. Those students who participated in three or more sessions that included educational activities had significantly higher scores on the scale relative to non-participants.

For the sample of youth from Boys & Girls Clubs referred to above under Reliability, higher scores on the scale were a marginally significant predictor of greater frequency and duration of attendance at the clubs over a 17-month period, controlling for students' demographic characteristics and measures of academic and behavioral risk factors. In longer-term follow-up analyses over a period of 29 months for 322 youth from this sample, youth with more frequent club attendance exhibited marginally significant increases in scores on the scale, controlling for youth demographic characteristics, club attended, measures of academic and delinquency risk and predictors of club participation from the baseline survey.

### Summary

Available evidence indicates only moderate reliability for the scale and is limited to a single sample with a relatively narrow age range. With regard to validity, available findings do not clearly establish the scale as a measure of school effort distinct from other constructs. To help address this issue, it would be useful to investigate the scale's association with well-validated measures of effort or investment in school work (convergent validity), as well as its relative degree of overlap with measures that target other youth attributes (discriminant validity). There is, however, evidence of the scale's ability to predict improvements in academic performance and to be an intervening variable in pathways linking measures of other youth attributes and experiences to outcomes in this area. It would be useful in future research to examine support for similar linkages of the scale with other types of youth outcomes. The scale also exhibits encouraging evidence of sensitivity to effects of OST program participation. Evaluation of OST program effects on the scale using an experimental design would add to findings in this area.

## Self-Efficacy

The Self-Efficacy scale from the *Beacons Youth Survey* is an eight-item self-report measure. Items on the scale are from the 17-item General Self-Efficacy Scale, with adaptations to make the wording appropriate for younger respondents. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. To date, there is no evidence pertaining to the scale's reliability. Evidence for the scale's validity is limited to moderate.

### Reliability

Evidence on the scale's reliability is not available.

### Validity

In path-modeling analyses conducted as part of a study of 838 students attending three middle schools that were implementing OST programming in the San Francisco Beacon Initiative, positive change on the scale between the fall of one school year and the spring of the following year was predictive of increased grades across the two school years both directly and via improvement in school effort. Change on the scale also was an intervening variable in pathways that linked measures of developmental experiences (e.g., support from non-familial adults) from the fall assessment to improvement in grades. In the same analyses, however, positive change on the scale was not directly or indirectly predictive of improvement on standardized test scores for reading or math and unexpectedly was found to be linked indirectly to increased absences. Further analyses compared the end-point spring scores on the scale for nine different groups of students who participated in varying numbers of program sessions (each of which lasted two to four months) that differed in their inclusion of educational and other activities with those for program non-participants, controlling for the baseline score and a range of additional variables. Students who participated in either one session with educational activities or two sessions without educational activities had significantly higher scores on the scale in comparison to non-participants. Scores for students in the other seven participation groups, however, did not differ from those of non-participants.

### Summary

There is a lack of information regarding the scale's reliability. With regard to validity, evidence is limited to results from a single study of youth participating in a particular type of OST program. These findings do not clearly establish the scale as a measure of self-efficacy beliefs distinct from other constructs. To help clarify this issue, it would be useful to investigate the scale's association with well-validated measures of the same construct (convergent validity), as well as its relative degree of overlap with measures that target youth attitudes or skills in other areas (discriminant validity). There is, however, evidence of the scale's ability to predict improvements in academic performance and to be part of pathways linking measures of other youth experiences and attributes to outcomes in this area. It would be useful in future research to examine support for similar linkages of the scale with other types of youth outcomes. Further investigation of the scale's sensitivity to effects of OST program participation also would be beneficial.

## Positive Reaction to Social Challenge

The Positive Reaction to Social Challenge scale from the *Beacons Youth Survey* is a six-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is limited and evidence for the scale's validity is moderate.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability within a sample of 246 participants in the Urban Corps Expansion Project who ranged in age from 15 to 28 years old. No data are available on test-retest reliability.

### Validity

In the same sample described under "Reliability," scores on the scale exhibited a significant positive association with a measure of the youth's constructive reaction to challenge as rated by education staff ( $r = .20$ ). Furthermore, in support of discriminant validity, the scale exhibited weaker and non-significant associations with measures of other, conceptually distinct constructs relating to performance and engagement that also were based on ratings from education staff. The scale's corresponding association with constructive reaction to challenge as rated by "crew leader" staff, however, was not significant, whereas it did exhibit a significant association with a measure of recognition within the domain of performance. The overall pattern of findings thus provides inconsistent support for the convergent and discriminant validity of the scale. For the same sample, scores on the scale exhibited theoretically expected associations with several youth-report measures of interpersonal support from peers, non-family adults and staff, as well as perceptions of competence in the area of education, but not with measures assessing support from family or perceptions of competence in the area of work.

In path modeling analyses conducted as part of a study of 838 students attending three middle schools that were implementing OST programming in the San Francisco Beacon Initiative, positive change on the scale between the fall of one school year and the spring of the following year was predictive of increased grades across the two school years indirectly via an association with improvement in school effort. Change on the scale also was an intervening variable in pathways that linked measures of developmental experiences (e.g., peer support) from the fall assessment to improvement in grades. In the same analyses, however, positive change on the scale was not directly or indirectly predictive of improvement on standardized test scores for reading or math and was unexpectedly linked indirectly to increased absences. Further analyses compared the end-point spring scores on the scale for nine different groups of students who participated in varying numbers of OST program sessions (each of which lasted two to four months) that differed in their inclusion of educational and other activities with those for program non-participants, controlling for the baseline score and a range of additional variables. Significant differences were not found between non-participants and students in any of the participation groups.

### Summary

Evidence of the scale's reliability is encouraging but is limited to data from a sample consisting of older adolescents and young adults. The scale's significant association with ratings of the same construct from adult observers provides noteworthy evidence of convergent validity. Available findings, however, are not consistent in this regard and do not clearly establish the scale as a measure of positive response to social challenge distinct from other constructs. To help clarify this issue, it would be useful to investigate the scale's association with well-validated measures of the same or similar constructs, as well as its relative degree of overlap with measures that target other youth attributes. The support found for the scale's role in pathways linking measures of other youth experiences and attributes to improvements in academic performance, although again not consistent across outcomes, is also noteworthy. It would be useful in future research to examine support for similar linkages of the scale with other types of youth outcomes. Further investigation of the scale's sensitivity to effects of OST program participation also would be beneficial.

## Leadership

The Leadership scale from the *Beacons Youth Survey* is an 11-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. To date, there is no evidence pertaining to the scale's reliability. Evidence for the scale's validity is none-to-limited.

### Reliability

Because scores on the scale are derived from reports about involvement in discrete activities that would not necessarily be expected to co-occur, the measure is formative rather than reflexive in orientation (see *Psychometrics: What Are They and Why are They Useful?* from the primary report *From Soft Skills to Hard Data* for discussion of this distinction). Consequently, test-retest reliability would be the most appropriate form of reliability to evaluate. Findings pertaining to this or other forms of reliability for the scale are not available.

### Validity

In path modeling analyses conducted as part of a study of 838 students attending three middle schools that were implementing OST programming in the San Francisco Beacon Initiative, scores on the scale from the fall of one school year were not predictive of change on measures of youth attributes (e.g., self-efficacy beliefs) between this time point and the spring of the following school year. The scale also was not linked to change in academic outcomes (grades, absences, math and reading test scores) across the two school years, either directly or indirectly. Further analyses compared the end-point spring scores on the scale for nine different groups of students who participated in varying numbers of program sessions (each of which lasted two to four months) that differed in their inclusion of educational and other activities with those for non-participants, controlling for the baseline score and a range of additional variables. Relative to non-participants, those students who participated in three or more sessions that did not include educational activities had marginally significantly higher scores on the scale, whereas marginally significantly lower scores were observed for those who participated in three or more sessions that included only educational activities. Significant differences were not found involving students in any of the other participation groups.

### Summary

There is a lack of information regarding the scale's reliability. With regard to validity, evidence is limited to results from a single study of youth participating in a particular type of OST program. These findings do not clearly establish the scale as a measure of leadership distinct from other constructs. To help clarify this issue, it would be useful to investigate the scale's association with well-validated measures of leadership (convergent validity), as well as its relative degree of overlap with measures that target youth attributes or abilities in other areas (discriminant validity). Available findings also do not provide support for theoretically expected associations with other measures or robust associations with OST program participation. Further examination of the scale's associations with relevant youth outcomes, as well as its sensitivity to OST program effects, thus would be beneficial.



## Time Spent in Challenging Learning Activity

The Time Spent in Challenging Learning Activity Leadership scale from the *Beacons Youth Survey* is a self-report measure. The measure asks youth to report the amount of time spent in each of eight different types of OST activities during a one-week period, as well as to rate how challenging they found each type of activity. A scale score is then derived by adding up the amount of time that was spent engaged in challenging activities. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. To date, there is no evidence pertaining to the scale's reliability. Evidence for the scale's validity is limited.

### Reliability

Because scores on the scale are derived from reports about involvement in discrete activities that would not necessarily be expected to co-occur, the measure is formative rather than reflexive (see Appendix 2 in the guide for discussion of this distinction). Consequently, test-retest reliability would be the most appropriate form of reliability to evaluate. Findings pertaining to this or other forms of reliability for the scale are not available.

### Validity

In path modeling analyses conducted as part of a study of 838 students attending three middle schools that were implementing OST programming in the San Francisco Beacon Initiative, scores on the scale from the fall of one school year were predictive of positive change between this time of assessment and the spring of the following school year on a measure of positive reaction to social challenge. Change on the scale was also linked indirectly, via its association with the measure of positive reaction to social challenge, to improvement on a measure of school effort and, in turn, school grades over the same time period. In the same analyses, however, the scale was not predictive of change on a measure of self-efficacy beliefs or improvement on standardized test scores in reading or math; higher scores also were found, unexpectedly, to be linked indirectly to increased absences. Further analyses compared the end-point spring scores on the scale for nine different groups of students who participated in varying numbers of program sessions (each of which lasted two to four months) that differed in their inclusion of educational and other activities with those for non-participants, controlling for the baseline score and a range of additional variables. Significant differences were not found between non-participants and students in any of the participation groups.

### Summary

There is a lack of information regarding the scale's reliability. With regard to validity, available findings do not clearly establish the scale as a measure of engagement in challenging learning activity distinct from other constructs. To help clarify this issue, it would be useful to investigate the scale's association with well-validated measures of the same construct (convergent validity), as well as its relative degree of overlap with measures that target youth attributes or skills in other areas (discriminant validity). The most promising validity evidence has linked positive change on the scale with improvements in academic performance via measures of theoretically relevant intervening constructs. It would be useful in future research to examine support for similar pathways involving other types of youth outcomes. Further investigation of the scale's sensitivity for detecting effects of OST program participation would also be beneficial.

# Technical Summaries of Scales Reviewed from the Social Skills Improvement System (SSIS) Rating Scales

## Communication

The Communication scales of the Teacher and Student Forms of the *SSIS Rating Scales* are seven- and six-item measures completed by a youth's teacher and the youth. For purposes of this guide, technical properties of the teacher- and youth-report versions of the scale were evaluated separately. In terms of the skill areas that are the focus of this guide, the content of the scales maps onto Communication.

### Teacher Form

Evidence for the reliability of the scale on the Teacher Form is moderate to substantial and evidence for the scale's validity is moderate.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from teachers of 550 youth ages 5-12 and 200 youth ages 13-18, including male and female youth within each age group. Test-retest reliability assessed over an average of 43 days for ratings provided by 144 teachers of youth ages 5-18 also was satisfactory ( $r = .76$ ), whereas interrater reliability for a sample of 54 pairs of teachers (or teacher and other school staff) also of youth ages 5-18 was moderate ( $r = .63$ ).

### Validity

In support of convergent validity, teacher ratings on the scale exhibited positive associations with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 723 youth ages 5-12 ( $r = .30$ ) and 289 youth ages 13-18 ( $r = .30$ ) and with youth ratings on the corresponding scale of the Student Form among 280 youth ages 8-12 ( $r = .28$ ) and 189 youth ages 13-18 ( $r = .23$ ). These associations were stronger than those that teacher ratings on the scale exhibited with other parent- and youth-report scales within the Social Skills portion of the instrument (mean  $r$ s of .27, .24, .25, and .15, respectively), but the margin of difference was noteworthy ( $> .05$ ) for only two of the four samples. In the national sample described under Reliability, teacher ratings on the scale also exhibited notably strong associations with teacher ratings on other subscales comprising the Social Skills scale ( $r$ s = .70 for youth ages 5-12 and 13-18). Overall, associations with ratings of social skills from teachers and other informants on the *SSIS Rating Scales* do not offer robust support for the scale's discriminant validity.

With further relevance to convergent and discriminant validity, for samples of 41 youth ages 5-12 and 16 youth ages 13-18, scores on the scale exhibited moderate to strong correlations (adjusted for restriction of range) with teacher ratings on the Functional Communication scale of the Behavioral Assessment Scale for Children-2 (*BASC-2*;  $r$ s of .66 and .83, respectively). These associations, however, were only modestly stronger than those that the scale exhibited with *BASC-2* scales focused on skills in conceptually distinct areas (Adaptability, Study Skills; mean  $r$ s of .63 and .75, respectively). Similarly, for a sample of 54 youth (44 of whom were ages 5-18), scores on the scale exhibited moderate correlations (adjusted for restriction of range) with both the Receptive and Expressive scales of the Communication composite of the Teacher Rating Form of the *Vineland Adaptive Behavior Scales, Second Edition (Vineland II-TRF)*;  $r$ s = .50). The scale's associations with scales comprising the Daily Living Skills composite (Personal, Academic, School/Community) of the *Vineland-II TRF* were of a comparable magnitude (mean  $r = .50$ ), which does not support discriminant validity, although, as would be expected, associations with the Motor Skills composite of the instrument were notably weaker (mean  $r = .25$ ).

In support of criterion validity, the scale exhibited correlations in expected directions with teacher ratings (national sample described under Reliability) on the Internalizing ( $r_s = -.33$  and  $-.43$  for youth ages 5-12 and 13-18, respectively), Externalizing ( $r_s = -.63$  and  $-.49$ ), Bullying ( $r_s = -.52$  and  $-.35$ ) and Academic Competence ( $r_s = .50$  and  $.53$ ) scales of the *SSIS Rating Scales*. Associations with parent and youth ratings on the same scales (excepting Academic Competence, which is not rated by these informants) from the Parent Form and Student Form samples described above were weaker, but still indicative of expected associations of noteworthy magnitude ( $r_s = .10$ ), with the exception of ratings obtained from older youth on the Student Form ( $r_s$  ranging from  $-.06$  to  $-.11$ ). For the samples described above in which teachers also rated youth on the *BASC-2*, the scale exhibited correlations (adjusted for restriction of range) in the expected negative direction with the following *BASC-2* scales: Aggression ( $r_s = -.41$  and  $-.36$  for youth ages 5-12 and 13-18, respectively), Conduct Problems ( $r_s = -.48$  and  $-.29$ ), Anxiety ( $r_s = -.13$  and  $-.33$ ), Depression ( $r_s = -.46$  and  $-.55$ ) and Somatization ( $r_s = -.11$  and  $-.39$ ). Similarly, for samples of 29 youth ages 5-12 and 10 youth ages 13-18 rated by teachers on the Elementary and Adolescent Versions of The *Walker-McConnell Scale of Social Competence and School Adjustment (SSCSA)*, respectively, ratings on the scale exhibited expected positive correlations with the School Adjustment scale of the *SSCSA* ( $r_s$  of  $.66$  and  $.71$ , adjusted for restriction of range). As also would be expected, for the national and Parent and Student Form samples youth with teacher, parent and self-report ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention and Autism Spectrum (teacher and parent ratings only) scales of the Problem Behaviors portion of the *SSIS Rating Scales* tended to receive lower ratings on the scale, as did those with more teacher-reported difficulties on the *BASC-2* Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales for the samples in which teachers also completed this measure.

### Summary

Evidence for the scale's reliability is generally encouraging. Validity evidence includes a supportive pattern of convergence with the corresponding subscale of the Parent and Student Forms of the *SSIS Rating Scales*, as well as with measures of communication skills as rated by teachers on other instruments. Available findings do not, however, clearly establish the scale as a measure of communication within social interactions distinct from other constructs. To address this concern, it would be useful to further investigate the scale's associations with well-validated measures of abilities in the area of communication (convergent validity) relative to measures that target skills in other areas (discriminant validity). The scale has exhibited expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well.

### Student Form

Evidence for the reliability of the scale on the Student Form is moderate to substantial and evidence for the scale's validity is moderate.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from 500 youth ages 8-12 and 300 youth ages 13-18, including male and female youth within each age group, as well as in a Spanish-speaking sample of 169 youth ages 5-12 and 65 youth ages 13-18 who completed the Spanish form of the scale. Test-retest reliability assessed over an average of 66 days for a sample of 127 youth ages 8-18 was moderate ( $r = .69$ ).

## Validity

In support of convergent validity, youth ratings on the scale exhibited positive associations with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 917 youth ages 8-12 ( $r = .22$ ) and 490 youth ages 13-18 ( $r = .25$ ) and with teacher ratings on the corresponding scale of the Teacher Form among 280 youth ages 8-12 ( $r = .28$ ) and 189 youth ages 13-18 ( $r = .23$ ). These associations were stronger than those that youth ratings on the scale exhibited with other parent- and teacher-report scales within the Social Skills portion of the instrument (mean  $r$ s of  $.14$ ,  $.23$ ,  $.23$ , and  $.08$ , respectively), but the margin of difference was noteworthy ( $> .05$ ) for only two of the four samples. Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* thus do not offer robust support for the scale's discriminant validity.

With further relevance to convergent and discriminant validity, for a sample of 34 youth ages 8-12 and 19 youth ages 13-18, scores on the scale exhibited moderate to strong correlations (adjusted for restriction of range) with youth ratings on the Interpersonal Relations scale of the Behavioral Assessment Scale for Children-2 (*BASC-2*;  $r$ s of  $.24$  and  $.58$ , respectively). These associations, however, were not consistently stronger than those that the scale exhibited with *BASC-2* scales focused on attitudes and skills in conceptually distinct areas (Attitude to School, Attitude to Teacher, Locus of Control, Self-Reliance; mean  $r$ s of  $.36$  and  $.48$ , respectively).

In support of criterion validity, youth ratings on the scale in most instances exhibited negative correlations with youth ratings (national sample described under Reliability) on the following scales of the Problem Behavior portion of the *SSIS Rating Scales*: Internalizing ( $r$ s =  $-.03$  and  $-.24$  for youth ages 8-12 and 13-18, respectively), Externalizing ( $r$ s =  $-.29$  and  $-.35$ ) and Bullying ( $r$ s =  $-.22$  and  $-.36$ ). Associations with parent and teacher ratings on the same scales from the Parent Form and Teacher Form samples described above were weaker, but still indicative of expected associations. For the samples described above in which youth also completed the *BASC-2* the scale exhibited correlations (adjusted for restriction of range) in expected directions with the following *BASC-2* scales: Social Stress ( $r$ s =  $-.30$  and  $-.55$  for youth ages 8-12 and 13-18, respectively), Anxiety ( $r$ s =  $-.27$  and  $-.40$ ), Depression ( $r$ s =  $-.38$  and  $-.14$ ), Sense of Inadequacy  $r$ s =  $-.29$  and  $-.39$ ) and Self-Esteem ( $r$ s =  $.28$  and  $.65$ ). As also would be expected, for the national and Parent and Teacher Form samples youth with self-report, parent, and teacher ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention and Autism Spectrum (teacher and parent ratings only) scales of the Problem Behaviors portion of the *SSIS Rating Scales* tended to provide lower ratings on the scale, as did those with more self-reported difficulties on the *BASC-2* Hyperactivity, Attention Problems and Atypicality scales for the samples that also completed this measure.

## Summary

Evidence for the scale's reliability is encouraging. Validity evidence includes a supportive pattern of convergence with the corresponding scale of the Parent and Teacher Forms of the instrument. Available findings do not, however, clearly establish the scale as a measure of communication within social interactions distinct from other constructs. To address this concern, it would be useful to further investigate the scale's associations with well-validated measures of abilities in the area of communication (convergent validity) relative to measures that target skills in other areas (discriminant validity). The scale has exhibited expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well.

## Assertion

The Assertion scales of the Teacher and Student Forms of the *SSIS Rating Scales* are seven-item measures completed by the youth's teacher and the youth. For purposes of this guide, technical properties of the teacher- and youth-report versions of the scale were evaluated separately. In terms of the skill areas that are the focus of this guide, the content of the scales maps onto Relationships and Collaboration.

### Teacher Form

To date, evidence for the reliability of the scale on the Teacher Form is moderate as is evidence for the scale's validity.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from teachers of 550 youth ages 5-12 and 200 youth ages 13-18, including male and female youth within each age group. Test-retest reliability assessed over an average of 43 days for ratings provided by 144 teachers of youth ages 3-18 was moderate ( $r = .68$ , adjusted for restriction of range), whereas inter-rater reliability for a sample of 54 pairs of teachers (or teacher and other school staff) also of youth ages 3-18 was low ( $r = .36$ , adjusted for restriction of range).

### Validity

Teachers ratings on the scale exhibited positive, but in most instances weak associations ( $r_s < .20$ ) with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 723 youth ages 5-12 ( $r = .23$ ) and 289 youth ages 13-18 ( $r = .16$ ) and with youth ratings on the corresponding scale of the Student Form among 280 youth ages 8-12 ( $r = .16$ ) and 189 youth ages 13-18 ( $r = .09$ ). These associations were stronger than those that teacher ratings on the scale exhibited with other parent- and youth-report scales within the Social Skills portion of the instrument (mean  $r_s$  of .19, .09, .11, and -.05, respectively), but the margin of difference was noteworthy ( $> .05$ ) for only two of the four samples. Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* do not offer robust support for the scale's convergent or discriminant validity.

With further relevance to convergent and discriminant validity, for samples of 41 youth ages 5-12 and 16 youth ages 13-18, scores on the scale exhibited moderate correlations (adjusted for restriction of range) with teacher ratings on the Social Skills scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)*;  $r_s$  of .56 and .40, respectively). These associations, however, were only marginally stronger than those the scale exhibited with ratings in the conceptually distinct area of Study Skills on the *BASC-2* ( $r_s$  of .52 and .37, respectively). For a sample of 54 youth (44 of whom were ages 5-18), however, scores on the scale did exhibit a notably stronger association with the Interpersonal Relationships scale of the Teacher Rating Form of the *Vineland Adaptive Behavior Scales, Second Edition* ( $r = .62$ , adjusted for restriction of range) than with scales comprising the Daily Living Skills (Personal, Academic, School/Community), Communication (Receptive, Expressive, Written) and Motor Skills (Gross, Fine) composites of the measure (mean  $r_s$  of .39, .21, and .21, respectively, adjusted for restriction of range).

With relevance to criterion validity, the scale exhibited correlations in expected directions with teacher ratings (national sample described under Reliability) for the Internalizing ( $r_s = -.25$  and  $-.29$  for youth ages 5-12 and 13-18, respectively) and Academic Competence ( $r_s = .32$ ) scales on the *SSIS Rating Scales*. As could be predicted for a scale intended to measure assertiveness, associations were not evident with the Externalizing and Bullying scales on the instrument. A similar pattern was evident when considering parent ratings (Parent Form sample described above) for the same scales (excepting Academic Competence which is not rated by parents or youth). For youth ratings (Student Form sample described above), however, the scale did not exhibit associations of note with either the Internalizing scale ( $r_s$  of  $-.01$  and  $-.05$ ) or the Externalizing and Bullying scales.

Similar to the pattern above, for the samples in which teachers also rated youth on the *BASC-2* the scale exhibited correlations (adjusted for restriction of range) in the expected negative direction with the Depression, Anxiety, and Somatization scales of the *BASC-2* (mean  $r_s = -.33$  and  $-.18$  for youth ages 5-12 and 13-18, respectively), but not with the Aggression and Conduct Problems scale (mean  $r_s = -.12$  and  $.04$ ). For samples of 29 youth ages 5-12 and 10 youth ages 13-18 rated by teachers on the Elementary and Adolescent Versions of The *Walker-McConnell Scale of Social Competence and School Adjustment (SSCSA)*, respectively, ratings on the scale exhibited expected positive associations with the School Adjustment scale of the *SSCSA* ( $r_s$  of  $.17$  and  $.57$ , adjusted for restriction of range). As also would be expected, for the national and Parent Form samples youth with teacher and parent ratings, respectively, indicative of more difficulties on the Autism Spectrum scale of the *SSIS Rating Scales* tended to receive lower ratings on the scale as did those with more teacher-reported difficulties on the *BASC-2* Attention Problems, Learning Problems, Withdrawal and Atypicality scales for the sample in which teachers also completed this measure.

## Summary

Evidence for the scale's reliability is generally encouraging, although test-retest reliability requires further investigation. With regard to validity, available findings do not provide clear support for the scale as a measure of assertiveness distinct from other components of social competence. To further examine this issue, it would be useful to investigate the scale in relation to well-validated measures of abilities in the same area from instruments other than the *SSIS Rating Scales*. The scale has generally exhibited a theoretically-expected pattern of association with indices of youth functioning in different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of relevant outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well.

## Student Form

To date, evidence for the reliability of the scale on the Student Form is moderate to substantial and evidence for the scale's validity is moderate.

## Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from 500 youth ages 8-12 and 300 youth ages 13-18, including male and female youth within each age group, as well as in a Spanish-speaking sample of 169 youth ages 5-12 and 65 youth ages 13-18 who completed the Spanish form of the scale. Test-retest reliability assessed over an average of 66 days for a sample of 127 youth ages 8-18 was moderate ( $r = .69$ ).

## Validity

Youth ratings on the scale exhibited positive, but in most instances weak associations ( $r_s < .20$ ) with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 917 youth ages 8-12 ( $r = .16$ ) and 490 youth ages 13-18 ( $r = .30$ ) and with teachers ratings on the corresponding scale of the Teacher Form among 280 youth ages 8-12 ( $r = .16$ ) and 189 youth ages 13-18 ( $r = .09$ ). These associations tended to be stronger than those that youth ratings on the scale exhibited with other parent- and teacher-report scales within the Social Skills portion of the instrument ( $r_s$  of  $.16$ ,  $.23$ ,  $.23$ , and  $.10$ , respectively), but the margin of difference was noteworthy ( $> .05$ ) for only two of the four samples. Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* thus do not offer robust support for the scale's convergent or discriminant validity.

With further relevance to convergent and discriminant validity, for a sample of 34 youth ages 8-12 and 19 youth ages 13-18, scores on the scale exhibited only weak to moderate positive correlations (adjusted for restriction of range) with youth ratings on the Interpersonal Relations scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)*;  $r_s$  of  $.09$  and  $.29$ , respectively). These associations, furthermore, were weaker than those that the scale exhibited with *BASC-2* scales focused on attitudes and skills in conceptually distinct areas (Attitude to School, Attitude to Teacher, Locus of Control, Self-Reliance; mean  $r_s$  of  $.24$  and  $.49$ , respectively).

With relevance to the scale's criterion validity, the scale consistently exhibited expected negative correlations of noteworthy magnitude ( $r_s \geq .10$ ) with ratings of youth ages 13-18, but not youth ages 8-12 (national sample described under Reliability), on the following scales of the Problem Behavior portion of the *SSIS Rating Scales*: Internalizing ( $r_s = .01$  and  $-.23$  for youth ages 8-12 and 13-18, respectively), Externalizing ( $r_s = -.19$  and  $-.30$ ) and Bullying ( $r_s = -.09$  and  $-.24$ ). Associations with parent and teacher ratings on the same scales from the Parent Form and Teacher Form samples described above were generally of noteworthy magnitude and in expected directions for each sample, with the exception of the Teacher Form sample for youth ages 13-18 (mean  $r = .00$ ). For the samples described above in which youth also completed the *BASC-2* the scale exhibited correlations (adjusted for restriction of range) in expected directions with the following *BASC-2* scales for youth ages 13-18, but not youth ages 8-12: Social Stress ( $r_s = -.03$  and  $-.30$  for youth ages 8-12 and 13-18, respectively), Anxiety ( $r_s = .24$  and  $-.24$ ), Depression ( $r_s = -.06$  and  $-.12$ ), Sense of Inadequacy ( $r_s = -.06$  and  $-.24$ ) and Self-Esteem ( $r_s = .05$  and  $.37$ ). For the national sample, youth with self-report ratings indicative of more difficulties on the Hyperactive/Inattention scale of the *SSIS Rating Scales* tended to provide lower ratings on the scale; a similar trend was not evident when considering teacher and parent ratings on the corresponding scale in the Teacher and Parent Form samples, although for these samples youth rated by teachers and parents as having more difficulties on the Autism Spectrum scale did tend to receive lower ratings on the scale. For the samples in which youth also completed the *BASC-2*, youth who reported more difficulties on the *BASC-2* Hyperactivity, Attention Problems and Atypicality scales tended to provide lower ratings on the scale.

### Summary

Evidence for the scale's reliability is encouraging. With regard to validity, available findings do not provide clear support for the scale as a measure of assertiveness distinct from other components of social competence. To further examine this issue, it would be useful to investigate the scale in relation to well-validated measures of abilities in the same area from instruments other than the *SSIS Rating Scales*. The scale has exhibited a somewhat inconsistent pattern of association with indices of youth functioning in different areas. In future work it would be helpful to also examine the scale's ability to contribute to prediction of relevant youth outcomes at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well. In view of the differences in findings for younger and older youth, attention to potential age differences in validity is also merited.

## Empathy

The Empathy scales of the Teacher and Student Forms of the *SSIS Rating Scales* are six-item measures completed by the youth's teacher and the youth. For purposes of this guide, technical properties of the teacher- and youth-report versions of the scale were evaluated separately. In terms of the skill areas that are the focus of this guide, the content of the scales maps onto Relationships and Collaboration.

### Teacher Form

Evidence for the reliability of the scale on the Teacher Form is moderate as is evidence for the scale's validity.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from teachers of 550 youth ages 5-12 and 200 youth ages 13-18, including male and female youth within each age group. Test-retest reliability assessed over an average of 43 days for ratings provided by 144 teachers of youth ages 3-18 also was satisfactory ( $r = .75$ , adjusted for restriction of range), whereas interrater reliability for a sample of 54 pairs of teachers (or teacher and other school staff) also of youth ages 3-18 was moderate ( $r = .57$ , adjusted for restriction of range).

### Validity

Teacher ratings on the scale were positively associated with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 723 youth ages 5-12 ( $r = .28$ ) and 289 youth ages 13-18 ( $r = .18$ ) and with youth ratings on the corresponding scale of the Student Form among 280 youth ages 8-12 ( $r = .29$ ), but not among 189 youth ages 13-18 ( $r = .00$ ). The former associations were for the most part only marginally stronger than those that teacher ratings on the scale exhibited with parent and youth ratings on other scales within the Social Skills portion of the instrument ( $r$ s of  $.24$ ,  $.15$ ,  $.18$ , respectively). Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* provide inconsistent support for the scale's convergent validity and limited support for the scale's discriminant validity.

With further relevance to convergent and discriminant validity, for samples of 41 youth ages 5-12 and 16 youth ages 13-18, scores on the scale exhibited moderate to strong correlations (adjusted for restriction of range) with teacher ratings on the Social Skills scale of the *Behavioral Assessment Scale for Children-2* (BASC-2;  $r$ s of  $.70$  and  $.64$ , respectively). These associations were stronger than those the scale exhibited with ratings for the conceptually distinct area of Study Skills on the BASC-2 ( $r$ s of  $.50$  and  $.56$ , respectively). Similarly, for a sample of 54 youth (44 of whom were ages 5-18), scores on the scale exhibited a stronger association with the Interpersonal Relationships scale of the Teacher Rating Form of the *Vineland Adaptive Behavior Scales, Second Edition* ( $r = .29$ ) than with scales comprising the Daily Living Skills (Personal, Academic, School/Community), Communication (Receptive, Expressive, Written) and Motor Skills (Gross, Fine) composites of the measure (mean  $r$ s of  $.12$ ,  $.22$ , and  $.12$ , respectively).



In support of criterion validity, the scale exhibited correlations in expected directions with teacher ratings (national sample described under Reliability) on the Internalizing ( $r_s = -.22$  and  $-.12$  for youth ages 5-12 and 13-18, respectively), Externalizing ( $r_s = -.52$  and  $-.28$ ), Bullying ( $r_s = -.47$  and  $-.23$ ) and Academic Competence ( $r_s = .29$  and  $.28$ ) scales of the *SSIS Rating Scales*. Associations with parent and youth ratings on the same scales (excepting Academic Competence, which is not rated by these informants) from the Parent Form and Student Form samples described above were weaker, but still indicative of expected associations. For the samples described above in which teachers also rated youth on the *BASC-2*, the scale also exhibited correlations (adjusted for restriction of range) in the expected negative direction with the following *BASC-2* scales: Aggression ( $r_s = -.32$  and  $-.27$  for youth ages 5-12 and 13-18, respectively), Conduct Problems ( $r_s = -.32$  and  $-.30$ ), Anxiety ( $r_s = -.28$  and  $-.57$ ), Depression ( $r_s = -.60$  and  $-.66$ ) and Somatization ( $r_s = -.32$  and  $-.48$ ). Similarly, for samples of 29 youth ages 5-12 and 10 youth ages 13-18 rated by teachers on the Elementary and Adolescent Versions of *The Walker-McConnell Scale of Social Competence and School Adjustment (SSCSA)*, respectively, ratings on the scale exhibited positive associations with the School Adjustment scale of the *SSCSA* ( $r_s$  of  $.30$  and  $.74$ , adjusted for restriction of range). As also would be expected, for the national and Parent and Student Form samples youth with teacher, parent and self-report ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention and Autism Spectrum (teacher and parent ratings only) scales of the *SSIS Rating Scales*, tended to receive lower ratings on the scale as did those with more teacher-reported difficulties on the *BASC-2* Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales for the samples in which teachers also completed this measure.

### Summary

Evidence for the scale's reliability is generally encouraging. With regard to validity, available findings do not provide strong support for the scale as a measure of empathy distinct from other components of social competence. To further address this issue, it would be useful to investigate the scale in relation to well-validated measures of abilities in the same area from instruments other than the *SSIS Rating Scales*. The scale has exhibited generally expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well. In view of the scale's lack of association with self-report ratings of empathy by older youth on the *SSIS Rating Scales*, attention to potential age differences in validity is also merited.

### Student Form

Evidence for the reliability of the scale on the Student Form is moderate to substantial and evidence for the scale's validity is moderate.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from 500 youth ages 8-12 and 300 youth ages 13-18, including male and female youth within each age group, as well as in a Spanish-speaking sample of 169 youth ages 5-12 and 65 youth ages 13-18 who completed the Spanish form of the scale. Test-retest reliability assessed over an average of 66 days for a sample of 127 youth ages 8-18 also was satisfactory ( $r = .76$ ).

### Validity

With relevance to convergent validity, youth ratings on the scale exhibited positive associations with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 917 youth ages 8-12 ( $r = .23$ ) and 490 youth ages 13-18 ( $r = .30$ ) and with teacher ratings on the corresponding scale of the Teacher Form among 280 youth ages 8-12 ( $r = .29$ ), but not among 189 youth ages 13-18 ( $r = .00$ ). The former associations were not consistently stronger than those that youth ratings on the scale exhibited with parent and teacher ratings on other scales within the Social Skills portion of the instrument ( $r_s$  of  $.23$ ,  $.18$ ,  $.29$ , respectively). Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* thus do not provide robust support for the scale's convergent or discriminant validity.

With further relevance to convergent and discriminant validity, for a sample of 34 youth ages 8-12 and 19 youth ages 13-18, scores on the scale exhibited a positive correlation (adjusted for restriction of range) with youth ratings on the Interpersonal Relations scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)* among the older sample ( $r = .48$ ), but not the younger sample ( $r = .02$ ). The former association, furthermore, was not stronger than those that the scale exhibited with *BASC-2* scales focused on attitudes and skills in conceptually distinct areas (Attitude to School, Attitude to Teacher, Locus of Control, Self-Reliance; mean  $r = .49$ ).

In support of criterion validity, youth ratings on the scale in most instances exhibited negative correlations of noteworthy magnitude ( $\geq .10$ ) with youth ratings (national sample described under Reliability) on the following scales of the Problem Behavior portion of the *SSIS Rating Scales*: Internalizing ( $r_s = .01$  and  $-.11$  for youth ages 8-12 and 13-18, respectively), Externalizing ( $r_s = -.27$  and  $-.29$ ) and Bullying ( $r_s = -.23$  and  $-.30$ ). Associations with parent and teacher ratings on the same scales from the Parent Form and Teacher Form samples described above were consistently indicative of expected associations and of noteworthy magnitude. For the samples described above in which youth also completed the *BASC-2*, the scale for the most part exhibited correlations (adjusted for restriction of range) of noteworthy magnitude in expected directions with the following *BASC-2* scales among youth ages 13-18, but not among youth ages 8-12: Social Stress ( $r_s = .05$  and  $-.47$  for youth ages 8-12 and 13-18, respectively), Anxiety ( $r_s = .12$  and  $-.34$ ), Depression ( $r_s = .00$  and  $-.05$ ), Sense of Inadequacy ( $r_s = -.04$  and  $-.50$ ) and Self-Esteem ( $r_s = .18$  and  $.46$ ). As would be expected, for the national and Parent and Teacher Form samples youth with self-report, parent and teacher ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention and Autism Spectrum (teacher and parent ratings only) scales of the *SSIS Rating Scales*, tended to provide lower ratings on the scale as did those with more self-reported difficulties on the *BASC-2* Hyperactivity, Attention Problems and Atypicality scales for the samples that also completed this measure.

## Summary

Evidence for the scale's reliability is encouraging. With regard to validity, available findings do not provide strong support for the scale as a measure of empathy distinct from other components of social competence. To further address this issue, it would be useful to investigate the scale in relation to well-validated measures of abilities in the same area from instruments other than the *SSIS Rating Scales*. The scale has exhibited generally expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well. In view of the scale's lack of association with teacher ratings of empathy for older youth on the *SSIS Rating Scales*, attention to potential age differences in validity is also merited.

## Engagement

The Engagement scales of the Teacher and Student Forms of the *SSIS Rating Scales* are seven-item measures completed by the youth's teacher and the youth. For purposes of this guide, technical properties of the teacher- and youth-report versions of the scale were evaluated separately. In terms of the skill areas that are the focus of this guide, the content of the scales maps onto Relationships and Collaboration.

## Teacher Form

Evidence for the reliability of the scale on the Teacher Form is moderate to substantial and evidence for the scale's validity is moderate.

## Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from teachers of 550 youth ages 5-12 and 200 youth ages 13-18, including male and female youth within each age group. Test-retest reliability assessed over an average of 43 days for ratings provided by 144 teachers of youth ages 3-18 also was satisfactory ( $r = .80$ , adjusted for restriction of range), whereas interrater reliability for a sample of 54 pairs of teachers (or teacher and other school staff) also of youth ages 3-18 was moderate ( $r = .69$ , adjusted for restriction of range).

## Validity

In support of convergent validity, teacher ratings on the scale were positively associated with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 723 youth ages 5-12 ( $r = .27$ ) and 289 youth ages 13-18 ( $r = .33$ ) and with youth ratings on the corresponding scale of the Student Form among 280 youth ages 8-12 ( $r = .17$ ) and 189 youth ages 13-18 ( $r = .31$ ). These associations exceed those that teacher ratings on the scale exhibited with other parent- and youth-report scales within the Social Skills portion of the instrument by a noteworthy margin ( $> .05$ ) for the samples of older youth (mean  $r$ s of  $.25$  and  $.14$ , respectively), but not the samples of younger youth (mean  $r$ s of  $.25$  and  $.18$ ). Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* do not provide robust support for the scale's discriminant validity.

With further relevance to convergent and discriminant validity, for samples of 41 youth ages 5-12 and 16 youth ages 13-18, scores on the scale exhibited strong correlations (adjusted for restriction of range) with teacher ratings on the Social Skills scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)* ( $r$ s of  $.75$  and  $.78$ , respectively). These associations were somewhat stronger than those the scale exhibited with ratings for the conceptually distinct area of Study Skills on the *BASC-2* ( $r$ s of  $.62$  and  $.75$ , respectively). Similarly, for a sample of 54 youth (44 of whom were ages 5-18), scores on the scale exhibited a stronger correlation (adjusted for restriction of range) with the Interpersonal Relationships scale of the Teacher Rating Form of the *Vineland Adaptive Behavior Scales, Second Edition* ( $r = .28$ ) than with scales comprising the Daily Living Skills (Personal, Academic, School/Community), Communication (Receptive, Expressive, Written) and Motor Skills (Gross, Fine) composites of the measure (mean  $r$ s of  $.18$ ,  $.05$ , and  $.02$ , respectively).

In support of criterion validity, the scale exhibited correlations in expected directions with teacher ratings (national sample described under Reliability) on the Internalizing ( $r$ s =  $-.40$  and  $-.55$  for youth ages 5-12 and 13-18, respectively), Externalizing ( $r$ s =  $-.33$  and  $-.27$ ), Bullying ( $r$ s =  $-.27$  and  $-.19$ ) and Academic Competence ( $r$ s =  $.41$  and  $.48$ ) scales of the *SSIS Rating Scales*. Associations with parent and youth ratings on the same scales (excepting Academic Competence, which is not rated by these informants) from the Parent Form and Student Form samples described above were weaker, but still indicative of expected associations, with the exception of ratings obtained from older youth on the Student Form ( $r$ s ranging from  $-.01$  to  $.05$ ). For the samples described above in which teachers also rated youth on the *BASC-2* the scale exhibited correlations (adjusted for restriction of range) in the expected negative direction with the following *BASC-2* scales: Aggression ( $r$ s =  $-.32$  and  $-.27$  for youth ages 5-12 and 13-18, respectively), Conduct Problems ( $r$ s =  $-.32$  and  $-.30$ ), Anxiety ( $r$ s =  $-.28$  and  $-.57$ ), Depression ( $r$ s =  $-.60$  and  $-.66$ ) and Somatization ( $r$ s =  $-.32$  and  $-.48$ ). For samples of 29 youth ages 5-12 and 10 youth ages 13-18 rated by teachers on the Elementary and Adolescent Versions of *The Walker-McConnell Scale of Social Competence and School Adjustment (SSCSA)*, respectively, ratings on the scale exhibited positive associations of varying magnitude with the School Adjustment scale of the *SSCSA* ( $r$ s of  $.09$  and  $.74$ , adjusted for restriction of range). As would be expected, for the national and Parent and Student Form samples youth with teacher, parent, and self-report ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention (teacher and parent ratings only) and Autism Spectrum scales of the *SSIS Rating Scales*, tended to receive lower ratings on the scale as did those with more teacher-reported difficulties on the *BASC-2* Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales for the samples in which teachers also completed this measure.

## Summary

Evidence for the scale's reliability is encouraging. Validity evidence includes a supportive pattern of convergence with the corresponding scale of the Parent and Student Forms of the *SSIS Rating Scales*. Available findings do not, however, clearly establish the scale as a measure of engagement distinct from other constructs. To address this concern, it would be useful to further investigate the scale's associations with well-validated measures of engagement (convergent validity) relative to measures that target skills in other areas (discriminant validity). The scale has exhibited expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well. In view of the scale's lack of association with self-report ratings of outcomes among older youth on the *SSIS Rating Scales*, attention to potential age differences in validity is also merited.

## Student Form

Evidence for the reliability of the scale on the Student Form is moderate to substantial as is evidence for the scale's validity.

## Reliability

The Student Form version of the scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from 500 youth ages 8-12 and 300 youth ages 13-18, including male and female youth within each age group, as well as in a Spanish-speaking sample of 169 youth ages 5-12 and 65 youth ages 13-18 who completed the Spanish form of the scale. Test-retest reliability assessed over an average of 66 days for a sample of 127 youth ages 8-18 was moderate ( $r = .58$ ).

## Validity

In support of convergent validity, youth ratings on the scale exhibited positive associations with parent ratings for the corresponding scale on the *SSIS Parent Form* among 917 youth ages 8-12 ( $r = .25$ ) and 490 youth ages 13-18 ( $r = .36$ ) and with teacher ratings on the corresponding scale of the *Teacher Form* among 280 youth ages 8-12 ( $r = .17$ ) and 189 youth ages 13-18 ( $r = .31$ ). In support of the scale's discriminant validity, these associations were consistently stronger than those that youth ratings on the scale exhibited with parent and teacher ratings on other scales within the Social Skills portion of the *SSIS* (mean  $r$ s of .14, .22, .11, and .09, respectively).

With further relevance to convergent and discriminant validity, for a sample of 34 youth ages 8-12 and 19 youth ages 13-18, scores on the scale exhibited positive correlations (adjusted for restriction of range) with youth ratings on the Interpersonal Relations scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)*;  $r$ s of .19 and .48, respectively). For the sample of older youth, but not younger youth, this association was stronger than those that the scale exhibited with *BASC-2* scales focused on attitudes and skills in conceptually distinct areas (Attitude to School, Attitude to Teacher, Locus of Control, Self-Reliance; mean  $r$ s of .30 and .32, for youth ages 5-12 and ages 13-18, respectively).

In support of criterion validity, youth ratings on the scale consistently exhibited negative correlations of noteworthy magnitude ( $\geq .10$ ) with youth ratings (national sample described under Reliability) on the following scales of the Problem Behavior portion of the *SSIS*: Internalizing ( $r$ s =  $-.15$  and  $-.24$  for youth ages 8-12 and 13-18, respectively), Externalizing ( $r$ s =  $-.21$  and  $-.20$ ) and Bullying ( $r$ s =  $-.14$  and  $-.21$ ).

Associations with parent and teacher ratings on the same SSIS scales from the Parent Form and Teacher Form samples described above also were indicative of expected associations and of noteworthy magnitude, with the exception of the scale's lack of associations with teacher ratings on the Externalizing and Bullying scales ( $r$ s ranging from .00 to -.09). For the samples described above in which youth also completed the BASC-2, the scale for the most part exhibited correlations (adjusted for restriction of range) of noteworthy magnitude in expected directions with the following BASC-2 scales: Social Stress ( $r$ s = -.06 and -.45 for youth ages 8-12 and 13-18, respectively), Anxiety ( $r$ s = .10 and -.19), Depression ( $r$ s = -.23 and -.13), Sense of Inadequacy ( $r$ s = -.22 and -.21) and Self-Esteem ( $r$ s = .42 and .38). As would be expected, for the national and Parent and Teacher SSIS samples youth with self-report, parent, and teacher ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention and Autism Spectrum (teacher and parent ratings only) scales of the SSIS tended to provide lower ratings on the scale as did those with more self-reported difficulties on the BASC-2 Hyperactivity, Attention Problems, and Atypicality scales for the samples that also completed this measure.

### Summary

Evidence for the scale's reliability is encouraging. Evidence for the scale's convergent and discriminant validity is also encouraging. It would be useful in future research to also examine the scale's associations with measures of engagement and skills in other areas that are derived from well-validated instruments other than the *SSIS Rating Scales*. The scale has exhibited generally expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well.

## Self-Control

The Self-Control scales of the Teacher and Student Forms of the *SSIS Rating Scales* are seven- and six-item measures completed by the youth's teacher and the youth, respectively. For purposes of this guide, technical properties of the teacher- and youth-report versions of the scale were evaluated separately. In terms of the skill areas that are the focus of this guide, the content of the scales maps onto Relationships and Collaboration.

### Teacher Form

Evidence for the reliability of the scale on the Teacher Form version of the scale is moderate to substantial and evidence for the scale's validity is moderate.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from teachers of 550 youth ages 5-12 and 200 youth ages 13-18, including male and female youth within each age group. Test-retest reliability assessed over an average of 43 days for ratings provided by 144 teachers of youth ages 3-18 also was satisfactory ( $r = .81$ , adjusted for restriction of range), whereas interrater reliability for a sample of 54 pairs of teachers (or teacher and other school staff) also of youth ages 3-18 was moderate ( $r = .62$ , adjusted for restriction of range).

### Validity

Teacher ratings on the scale were positively associated with parent ratings for the same scale on the Parent Form of the *SSIS Rating Scales* among 723 youth ages 5-12 ( $r = .29$ ) and 289 youth ages 13-18 ( $r = .30$ ) and with youth ratings on the corresponding scale of the Student Form among 280 youth ages 8-12 ( $r = .30$ ), but not among 189 youth ages 13-18 ( $r = .04$ ). The scale's associations with corresponding parent ratings notably exceeded those that it exhibited with other parent-report scales within the Social Skills portion of the instrument (mean  $r$ s of  $.21$  for both age groups), whereas its association with corresponding youth ratings for the age 8-12 sample only marginally exceeded those that it exhibited with other youth-report scales (mean  $r = .27$ ). Overall, associations with ratings of social skills from other informants on the *SSIS Rating Scales* thus do not provide consistent support for the scale's convergent or discriminant validity.

With further relevance to convergent and discriminant validity, for samples of 41 youth ages 5-12 and 16 youth ages 13-18, scores on the scale exhibited strong correlations (adjusted for restriction of range) with teacher ratings on the Social Skills scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)*;  $r$ s of  $.78$  and  $.73$ , respectively). These associations do not, however, consistently exceed those the scale exhibited with ratings in the conceptually distinct areas of Leadership and Study Skills on the *BASC-2* (mean  $r$ s of  $.73$  and  $.81$ , respectively). Similarly, for a sample of 54 youth (44 of whom were ages 5-18), scores on the scale exhibited a notably stronger correlation (adjusted for restriction of range) with the Interpersonal Relationships scale of the Teacher Rating Form of the *Vineland Adaptive Behavior Scales, Second Edition (Vineland-II TRF)*;  $r = .41$ ) than with scales comprising the Communication (Receptive, Expressive, Written) and Motor Skills (Gross, Fine) composites of the measure (mean  $r$ s of  $.19$  and  $.10$ , respectively), but not in comparison to those comprising the Daily Living Skills (Personal, Domestic, Community) composite (mean  $r = .40$ ). Finally, for a sample of 10 youth ages 13-18 also rated by teachers on the Adolescent Versions of *The Walker-McConnell Scale of Social Competence and School Adjustment (SSCSA)*, scores on the scale exhibited a strong correlation (adjusted for restriction of range) with the Self-Control scale of the *SSCSA* ( $r = .84$ ), which was somewhat stronger than its corresponding association with the Empathy scale on the same instrument ( $r = .71$ ).

In support of criterion validity, the scale exhibited correlations in expected directions with teacher ratings (national sample described under Reliability) on the Internalizing ( $r_s = -.36$  and  $-.25$  for youth ages 5-12 and 13-18, respectively), Externalizing ( $r_s = -.66$  and  $-.49$ ), Bullying ( $r_s = -.54$  and  $-.39$ ), and Academic Competence ( $r_s = .32$  and  $.39$ ) scales of the *SSIS Rating Scales*. Associations with parent and youth ratings on the same scales (excepting Academic Competence, which is not rated by these informants) from the Parent Form and Student Form samples described above were weaker, but still indicative of expected associations, with the exception of ratings obtained from older youth on the Student Form ( $r_s$  ranging from  $-.04$  to  $-.10$ ). For the samples described above in which teachers also rated youth on the BASC-2, the scale generally exhibited correlations (adjusted for restriction of range) in the expected negative direction with the following BASC-2 scales: Aggression ( $r_s = -.69$  and  $-.42$  for youth ages 5-12 and 13-18, respectively), Conduct Problems ( $r_s = -.65$  and  $-.49$ ), Anxiety ( $r_s = -.15$  and  $-.51$ ), Depression ( $r_s = -.62$  and  $-.61$ ), and Somatization ( $r_s = -.14$  and  $-.43$ ). For a sample of 29 youth ages 5-12 rated by teachers on the Elementary version of the SSCSA and the previously described sample of 10 youth ages 13-18 rated by teachers on the Adolescent version of the instrument, ratings on the scale similarly exhibited positive associations with the School Adjustment scale of the measure ( $r_s$  of  $.38$  and  $.85$ , respectively, adjusted for restriction of range). As would be expected, for the national and Parent and Student Form samples youth with teacher, parent, and self-report ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention (teacher and parent ratings only) and Autism Spectrum scales of the *SSIS Rating Scales*, tended to receive lower ratings on the scale as did those with more teacher-reported difficulties on the BASC-2 Hyperactivity, Attention Problems, Learning Problems, Withdrawal and Atypicality scales for the samples in which teachers also completed this measure.

### Summary

Evidence for the scale's reliability is encouraging. With regard to validity, available findings do not provide strong support for the scale as a measure of self-control distinct from other components of social competence. To further address this issue, it would be useful to investigate the scale in relation to well-validated measures of abilities in the same area from instruments other than the *SSIS Rating Scales*. The scale has exhibited expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well. In view of the scale's lack of association with self-report ratings of outcomes among older youth on the *SSIS Rating Scales*, attention to potential age differences in validity is also merited.

### Student Form

Evidence for the reliability of the scale on the Student Form is moderate to substantial and evidence for validity is moderate.

### Reliability

This scale demonstrated satisfactory internal consistency reliability in a national sample of ratings from 500 youth ages 8-12 and 300 youth ages 13-18, including male and female youth within each age group, as well as in a Spanish-speaking sample of 169 youth ages 5-12 and 65 youth ages 13-18 who completed the Spanish form of the scale. Test-retest reliability assessed over an average of 66 days for a sample of 127 youth ages 8-18 was moderate ( $r = .62$ ).

### Validity

With relevance to convergent validity, youth ratings on the scale exhibited positive associations with parent ratings for the corresponding scale on the Parent Form of the *SSIS Rating Scales* among 917 youth ages 8-12 ( $r = .22$ ) and 490 youth ages 13-18 ( $r = .28$ ) and with teacher ratings on the corresponding scale of the Teacher Form among 280 youth ages 8-12 ( $r = .30$ ), but not among 189 youth ages 13-18 ( $r = .04$ ). In support of discriminant validity, the former associations were stronger than those that youth ratings on the scale exhibited with parent and teacher ratings on other scales within the Social Skills portion of the instrument (mean  $r_s$  of  $.17$ ,  $.22$ ,  $.21$ , respectively).

With further relevance to convergent and discriminant validity, for a sample of 34 youth ages 8-12 and 19 youth ages 13-18, scores on the scale exhibited a positive correlation of noteworthy magnitude (adjusted for restriction of range) with youth ratings on the Interpersonal Relations scale of the *Behavioral Assessment Scale for Children-2 (BASC-2)* among the older sample ( $r = .30$ ), but not the younger sample ( $r = .02$ ). The former association, furthermore, was weaker than those that the scale exhibited with *BASC-2* scales focused on attitudes and skills in conceptually distinct areas (Attitude to School, Attitude to Teacher, Locus of Control, Self-Reliance; mean  $r = .54$ ).

In support of criterion validity, youth ratings on the scale consistently exhibited negative correlations with youth ratings (national sample described under Reliability) on the following scales of the Problem Behavior portion of the *SSIS Rating Scales*: Internalizing ( $r_s = -.10$  and  $-.32$  for youth ages 8-12 and 13-18, respectively), Externalizing ( $r_s = -.31$  and  $-.49$ ) and Bullying ( $r_s = -.20$  and  $-.39$ ). Associations with parent and teacher ratings on the same scales from the Parent Form and Teacher Form samples described above also were generally indicative of expected associations. For the samples described above in which youth also completed the *BASC-2* the scale similarly exhibited correlations (adjusted for restriction of range) in expected directions with the following *BASC-2* scales: Social Stress ( $r_s = -.26$  and  $-.53$  for youth ages 8-12 and 13-18, respectively), Anxiety ( $r_s = -.23$  and  $-.40$ ), Depression ( $r_s = -.24$  and  $-.63$ ), Sense of Inadequacy ( $r_s = -.09$  and  $-.46$ ) and Self-Esteem ( $r_s = .09$  and  $.57$ ). As would be expected, for the national and Parent and Teacher Form samples youth with self-report, parent, and teacher ratings, respectively, indicative of more difficulties on the Hyperactive/Inattention and Autism Spectrum (teacher and parent ratings only) scales of the *SSIS Rating Scales*, tended to provide lower ratings on the scale as did those with more self-reported difficulties on the *BASC-2* Hyperactivity, Attention Problems and Atypicality scales for the samples that also completed this measure.

### Summary

Evidence for the scale's reliability is encouraging. With regard to validity, available findings do not provide clear support for the scale as a measure of self-control in social interactions distinct from other components of social competence. To further examine this issue, it would be useful to investigate the scale in relation to well-validated measures of abilities in the same area from instruments other than the *SSIS Rating Scales*. The scale has exhibited generally expected associations with indices of youth functioning in a range of different areas. It would be helpful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Investigation of the scale's sensitivity for detecting expected effects on youth participating in OST programs would be valuable as well. In view of the scale's lack of association with teacher ratings of self-control for older youth on the *SSIS Rating Scales*, attention to potential age differences in validity is also merited.



# Technical Summaries of Scales Reviewed from the Survey of Academic and Youth Outcomes (SAYO)

## Behavior in the Classroom

The Behavior in the Classroom scale of the SAYO is completed by the youth's teacher and consists of four items. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is limited to moderate to substantial as is evidence for the scale's validity.

### Reliability

The scale has demonstrated satisfactory internal consistency in a large sample of youth in grades K-12 participating in OST programs in 27 school districts. Internal consistency reliability estimates were satisfactory for male and female youth and for each of three different age groups ranging from youngest (age or grade levels not defined) to high school students. No data are available on interrater or test-retest reliability.

### Validity

With relevance to convergent validity, in analyses of ratings provided by teachers of 251 youth from 17 CCLC program sites in Massachusetts, scores on the scale exhibited a significant and positive correlation with program staff ratings of youth behavior in the OST program on the Behavior in the Program scale of the SAYO-S ( $r = .58$ ). The scale's associations with program staff ratings of youth on SAYO scales assessing problem-solving, engagement and relations with adults, initiative, and communication, however, were similarly strong (mean  $r = .56$ ). Overall, findings did not offer robust support for the scale's discriminant validity. In support of criterion validity, for the same sample ratings on the scale exhibited significant correlations with ratings that teachers provided of the quality of a youth's work on homework assignments ( $r = .79$ ) and academic performance ( $r$ s ranging from  $.34$  to  $.48$ ) on the SAYO-T. Significant correlations were also found with youths' self-reports on the SAYO of their academic competence in three of four subject areas (significant  $r$ s ranging from  $.19$  to  $.28$ ). In further findings based on this sample, teacher ratings on the scale exhibited significant associations with SAYO scales that ask youth to retrospectively report their personal/social ( $r = .23$ ) and academic ( $r = .26$ ) gains associated with their OST program participation.

In other research with youth participating in 78 OST programs in Massachusetts, a measure of program attendance was not found to be related significantly to change in scores on the scale over the course of a school year. In addition, change in scores on the scale did not vary significantly across program sites, despite evidence that the sites varied significantly in terms of their quality. Teacher ratings on the scale did, however, improve significantly between fall and spring assessments for a sample of 1,060 youth participating in OST programs.

## Summary

This scale shows evidence of being reliable when used by teachers to rate youth in different age groups, as well as male and female youth. The finding of a substantial association with program staff ratings of youth behavior provides encouraging evidence of convergent validity. Available evidence does not, however, clearly establish the scale as a measure of classroom behavior distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures that have a similar focus (convergent validity) relative to those that target skills or behaviors in other areas (discriminant validity). In view of the finding of a notably strong association with teacher ratings of the quality of homework assignments completed, the extent to which the scale is sufficiently specific to be regarded as a measure of the appropriateness of a youth's classroom behavior distinct from constructs relating to academic performance would benefit from additional investigation. The scale's concurrent associations with teacher- and youth-report indices pertaining to academic competence do, however, provide encouraging evidence of criterion validity. It would be useful to build on these findings by examining the scale's associations with similar outcomes at later points in youths' schooling or development (predictive validity). Evidence of the scale's associations with youth participation in OST programs is mixed. Further investigation of the scale's sensitivity to effects of OST program participation using a quasi-experimental or randomized control design would be valuable. Currently, all available findings involve other scales within the SAYO assessment system. It would be helpful for future research on the scale to include well-validated measures from other sources.

## Initiative

The Initiative scale of the SAYO is designed to be completed by the youth's teacher or OST program staff. The teacher and OST program staff versions of the scale each consists of five items. They differ somewhat in item wording and content but were evaluated together for purposes of this compendium. In terms of the skill areas that are the focus of this guide, the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

Ratings of OST program staff and teachers on the scale demonstrated satisfactory internal consistency reliability in two samples of youth participating in OST programs in Massachusetts and Pennsylvania (staff ratings only). Estimates of internal reliability for male and female youth, for elementary, middle and high school youth, and for Black, Hispanic and White youth in these samples were satisfactory. Data on interrater or test-retest reliability are not available for either type of rater.

### Validity

In support of convergent validity, in analyses of data for a sample of 224 youth from 17 CCLC program sites in Massachusetts, a substantial positive correlation was found between OST program staff and teacher ratings on the scale ( $r = .54$ ). Correlations of ratings on the scale for each type of informant with ratings of the other informant on the teacher and staff versions of the SAYO (Problem Solving, Engagement, Communication Skills, Behavior, Adult Relations) were also reported. For both OST program staff and teacher ratings only two of the cross-construct correlations were lower than the convergent correlation reported above. Findings thus did not provide robust support for the scale's discriminant validity. In further findings for this sample relevant to convergent validity, staff ratings on the scale, but not teacher ratings, exhibited a significant positive association with youth ratings on the Future Planning-My Actions scale of SAYO-Y ( $r$ s of  $.37$  and  $.00$ , respectively), which asks youth about their initiative and future-orientation in relation to school work. Staff ratings were also associated significantly in the expected positive direction with SAYO-Y scales that ask youth about their experiences in OST programs ( $r$ s =  $.14$  to  $.24$ ). Staff ratings were associated positively as well with the SAYO-Y scale that asks youth to retrospectively report personal/social gains associated with their program participation ( $r = .18$ ), although not with a corresponding scale focused on academic change; neither correlation was significant for teacher ratings on the scale. In support of criterion validity, ratings of staff on the scale exhibited significant and strong correlations with teacher ratings of the quality of the youth's homework assignments ( $r = .83$ ) and academic performance ( $r$ s ranging from  $.53$  to  $.67$ ) on the SAYO-T. In general, however, neither teacher nor staff ratings on the scale exhibited significant associations with scales on the SAYO-Y that assess youth reports of their sense of academic competence. In path modeling analyses with a sample of 442 youth from the same study, teacher ratings of the scale were an intervening variable in several significant and theoretically interpretable indirect pathways that linked youth reports of their OST program experiences with teacher ratings of their academic performance on the SAYO-T. Staff ratings on the scale also were linked to teacher ratings of academic performance via pathways that involved staff ratings of performance on homework as an intervening variable.

In two other samples, teacher and staff ratings on the measure both exhibited expected improvements over the course of a year in which a youth was participating in an OST program. In other research with youth participating in 78 OST programs in Massachusetts, youth attending programs objectively assessed as being higher on youth engagement and family relations dimensions of quality exhibited significantly more favorable change on the measure for staff ratings ( $r$ s =  $.32$  and  $.26$ , respectively). A corresponding association for a global measure of program quality was somewhat weaker and non-significant ( $r = .20$ ). A measure of program attendance, furthermore, was not related to change in either program staff or teacher ratings on the scale.

## Summary

This scale shows evidence of being reliable when used by teachers or OST program staff to rate demographically varied groups of youth. The finding of a substantial association between teacher and program staff ratings of youth on the scale provides encouraging evidence of convergent validity. Available evidence does not, however, clearly establish the scale as a measure of initiative distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of this construct relative to measures of youth skills or behaviors in other areas. In view of the notably strong association found between teacher ratings on the scale and their ratings of the quality of the youth's homework assignments, the extent to which it is sufficiently specific to be regarded as a measure of initiative distinct from constructs relating to performance on school work is one issue that would be useful to investigate. The scale's concurrent associations with teacher- and youth-report indices pertaining to academic competence do, however, provide encouraging evidence of criterion validity. It would be useful to build on these findings by examining the scale's association with outcomes in other areas, as well as those assessed at later points in youths' schooling or development (predictive validity). There is also evidence of expected associations of the scale with OST program participation, most notably greater improvement over time in scores for youth attending programs with greater objectively assessed quality. It would be useful to add to findings in this area by investigating the scale's sensitivity to effects of program participation; using a quasi-experimental or randomized control design would be valuable. Currently, all available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

## Relations with Adults

The Relations with Adults scale of the SAYO is designed to be completed by the youth's teacher or OST program staff. The teacher and OST program staff versions consist of four and five items, respectively. Although they differ somewhat in item wording and content, for purposes of this guide they were evaluated together. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

Ratings of OST program staff on this scale demonstrated satisfactory internal consistency reliability for two samples of youth participating in OST programs in Massachusetts and Pennsylvania. Estimates of internal reliability for male and female youth, for elementary, middle and high school youth, and for Black, Hispanic and White youth in these samples were satisfactory. Satisfactory internal consistency reliability for teacher ratings has also been reported for a sample of 221 youth. Data on interrater or test-retest reliability are not available for either type of rater.

### Validity

In support of convergent validity, a significant correlation ( $r = .40$ ) was found between ratings of OST program staff and teachers on the scale for a sample of 359 youth from 17 CCLC program sites in Massachusetts. Correlations of ratings on the scale for each type of informant with ratings of the other informant on five other scales from teacher and staff versions of the SAYO (Problem Solving, Engagement, Initiative, Behavior, Communication Skills) also were reported. In support of discriminant validity, for OST program staff ratings all but one of the cross-construct correlations was lower than the convergent correlation reported above. For teacher ratings, however, no correlations met this criterion. For this same sample, a SAYO scale in which youth retrospectively report personal/social gains associated with program participation exhibited significant associations with both staff and teacher ratings on the scale ( $r_s = .16$ ); retrospective youth reports of academic change were associated significantly with teacher ratings ( $r = .21$ ) but not staff ratings. With relevance to criterion validity, ratings of both teachers and staff on the scale from the same sample exhibited significant correlations in expected positive directions with teacher ratings on the SAYO-T of both the quality of work on homework assignments ( $r_s$  of  $.62$  and  $.34$  for ratings of teachers and staff, respectively) and academic performance ( $r_s$  ranging from  $.30$  to  $.39$  and from  $.24$  to  $.34$ , respectively). Teacher ratings on the scale also exhibited significant positive associations with SAYO-Y scales asking youth to report on their sense of academic competence, although for OST program staff ratings this was the case for only two of the five scales involved.

In two other samples, staff ratings on the scale improved significantly over the course of a year in which a student was participating in an OST program. In further research with youth participating in 78 OST programs in Massachusetts, youth attending programs that were objectively assessed as higher on a family relations dimension of quality exhibited significantly more favorable change on the scale for staff ratings ( $r = .25$ ). Corresponding associations with other dimensions of program quality were not significant ( $r_s$  ranging from  $.02$  to  $.23$ ). The association of a global program quality measure with change on the scale was similarly positive, but not significant ( $r = .23$ ). A measure of program attendance, furthermore, was related in an unexpected negative direction to change in staff ratings from a pretest assessment on the scale and there was no association in relation to change in teacher ratings. Finally, in path modeling analyses conducted with a sample of 442 youth in grades 4 through 8 who were participating in the 21st Century Community Learning Center (CCLC) program sites in Massachusetts, teacher ratings on the scale were not involved in any of the significant indirect effects that linked youth reports of their OST program experiences on the SAYO-Y to their academic performance as rated by teachers on the SAYO-T.

## Summary

This scale shows evidence of being reliable when used by teachers or OST program staff to rate demographically varied groups of youth. The finding of a substantial association between teacher and program staff ratings of youth on the scale provides encouraging evidence of convergent validity. Available evidence does not, however, clearly establish the scale as a measure of the quality of the youth's relationships with adults distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of this construct relative to measures that target behavior or skills in other areas. It would be useful in future research to examine the extent to which the scale taps specifically into skills for relating to adults as distinct from peers. In view of the notably strong association the scale has exhibited with teacher ratings of the quality of homework assignments completed, the extent to which the scale is sufficiently specific to be regarded as a measure of the youth's relationships with adults as distinct from constructs relating to school work also would benefit from further investigation. The scale's concurrent associations with teacher- and youth-report indices pertaining to academic competence do, however, provide encouraging evidence of criterion validity. It would be useful to build on these findings by examining the scale's association with outcomes in other areas, as well as those assessed at later points in youths' schooling or development (predictive validity). There is also evidence of expected associations of the scale with OST program participation, although findings are mixed. Further investigation of the scale's sensitivity to effects of program participation using a quasi-experimental or randomized control design would be valuable. Currently, all available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

## Relations with Peers

The Relations with Peers scale of the SAYO is designed to be completed by the youth's teacher or OST program staff. The teacher and staff versions of the scale consist of three and four items, respectively. They differ somewhat in item wording and content, but for purposes of this guide were evaluated together. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

Ratings of OST program staff on this scale demonstrated satisfactory internal consistency reliability for two samples of youth participating in OST programs in Massachusetts and Pennsylvania. Estimates of internal reliability for male and female youth, for elementary, middle and high school youth, and for Black, Hispanic and White youth in these samples were satisfactory. Satisfactory internal consistency reliability for teacher ratings also has been reported for a sample of 221 youth. Data on interrater or test-retest reliability data are not available for either type of rater.

### Validity

In support of convergent validity, in analyses of data for 271 youth from 17 CCLC program sites across five Massachusetts school districts, OST program staff ratings on the scale exhibited a significant positive correlation with youth ratings on the Peer Relations scale of the SAYO-Y ( $r = .22$ ). For this same sample, two SAYO-Y scales that ask youth about their experiences in OST programs (supportive adults, enjoyment/engagement) exhibited significant associations in the expected positive direction with staff ratings on the scale ( $r_s = .13$ ). Associations with three other scales, however, including one that is focused on the supportiveness of the program's social environment, were non-significant ( $r_s = -.01$  to  $.09$ ). In support of criterion validity, for the same sample ratings of OST program staff on the scale exhibited significant positive correlations with teacher ratings of the youth's quality of work on homework assignments and academic performance on the SAYO-T.

In other findings relevant to validity, in two distinct samples staff ratings on the scale improved significantly over the course of a year in which a student was participating in an OST program. In another sample involving youth attending 78 OST programs in Massachusetts, those attending programs that were objectively assessed as being higher on a youth engagement dimension of quality exhibited significantly more favorable change on the scale for staff ratings ( $r = .43$ ). A corresponding pattern was not evident for other dimensions of program quality ( $r_s = -.08$  to  $.19$ ), although a significant association was found when examining a global measure of program quality ( $r = .25$ ). A measure of program attendance was related in an unexpected negative direction to change in staff ratings on the scale and there was no association in relation to change in teacher ratings.

### Summary

This scale shows evidence of being reliable when used by teachers or OST program staff to rate demographically varied groups of youth. The association of the scale with youth self-report ratings of their peer relations, although limited in strength, provides encouraging evidence of convergent validity. Available evidence does not, however, establish the scale as a measure of social competence distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of social competence (convergent validity) relative to measures that target skills or behaviors in other areas (discriminant validity). The scale has demonstrated concurrent associations with indices of youth school work and academic competence. It would be useful to build on these findings by examining the scale's association with outcomes in other areas, as well as those assessed at later points in youths' schooling or development (predictive validity). There is also evidence of expected associations with OST program participation for the scale, most notably greater improvement over time in scores for youth attending programs with greater objectively assessed quality. It would be useful to add to findings in this area by investigating the scale's sensitivity to effects of OST program participation using a quasi-experimental or randomized control design. All available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

## Problem Solving Skills

The Problem Solving Skills scale of the SAYO is designed to be completed by the youth's teacher or OST program staff. The teacher and staff versions of the scale consist of five and three items, respectively. They differ somewhat in item wording and content but were evaluated together for purposes of this compendium. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Critical Thinking and Decision-Making. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

Ratings of OST program staff and teachers on the scale demonstrated satisfactory internal consistency reliability in two samples of youth participating in OST programs in Massachusetts and Pennsylvania (staff ratings only). Estimates of internal reliability for male and female youth, for elementary, middle and high school youth, and for Black, Hispanic and White youth in these samples were satisfactory. Data on interrater or test-retest reliability are not available for either type of rater. Data are not available on test-retest reliability for either type of rater.

### Validity

In support of convergent validity, in analyses of data for a sample of 89 youth from 17 CCLC program sites in Massachusetts a substantial positive correlation was found between OST program staff and teacher ratings on the scale ( $r = .64$ ). Correlations of ratings on the scale for each type of informant with ratings of the other informant on the teacher and staff versions of the SAYO (Initiative, Engagement, Communication Skills, Behavior, Adult Relations) were also reported. In support of discriminant validity, for both OST program staff and teacher ratings all but one of the cross-construct correlations was lower than the convergent validity correlation reported above. The margin of this difference, however, tended to be small. For this same sample, scales on the SAYO-Y that ask youth about their experiences in OST programs were not significantly associated with staff ratings on the scale; SAYO-Y scales that ask youth to retrospectively report academic or personal/social change associated with their OST program participation also did not exhibit significant associations with staff or teacher ratings on the scale. In support of criterion validity, ratings of staff on the scale did exhibit significant and strong correlations with teacher ratings of the quality of the youth's homework assignments ( $r = .69$ ) and academic performance ( $r$ s ranging from  $.37$  to  $.48$ ) on the SAYO-T. For the most part, however, neither teacher nor staff ratings on the scale exhibited significant associations with youth self-reports on the SAYO-Y of their academic competence. Finally, in path modeling analyses with a sample of 442 youth from the same study, teacher ratings on the scale were an intervening variable in several significant and theoretically interpretable indirect pathways that linked youth reports of their OST program experiences with teacher-rated academic outcomes. Staff ratings on the scale also were linked to teacher ratings of academic performance via pathways that involved staff ratings of the quality of the youth's work on homework assignments as an intervening variable.

In other research with youth participating in 78 OST programs in Massachusetts, a measure of program attendance was not found to be significantly related to change in teacher ratings on the scale over the course of a school year. In addition, change in teacher ratings on the scale did not vary significantly across program sites, despite evidence that the sites varied significantly in terms of their quality. Finally, in a separate sample, teacher ratings on the scale improved significantly over the course of a year in which a student participated in an OST program.



## Summary

This scale shows evidence of being reliable when used by teachers or OST program staff to rate demographically varied groups of youth. The finding of a substantial association between teacher and program staff ratings of youth on the scale provides encouraging evidence of convergent validity. Available evidence does not, however, clearly establish the scale as a measure of problem solving distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of problem solving relative to measures of youth skills in other areas. In view of the notably strong association found between teacher ratings on the scale and their ratings of the quality of the youth's homework assignments, the extent to which it is sufficiently specific to be regarded as a measure of problem-solving ability distinct from constructs relating to performance on school work is one issue that would be useful to investigate. The scale's concurrent associations with teacher reports of academic performance do, however, provide encouraging evidence of criterion validity. It would be useful to build on these findings by examining the scale's association with outcomes in other areas, as well as those assessed at later points in youths' schooling or development (predictive validity). Evidence of expected associations of the scale with OST program participation is limited. Further investigation of the scale's sensitivity to effects of OST program participation using a quasi-experimental or randomized control design would be valuable. All available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

## Communication Skills

The Communication Skills scale of the SAYO is designed to be completed by the youth's teacher or OST program staff. The teacher and staff versions of the scale consist of five and four items, respectively. They differ somewhat in item wording and content but were evaluated together for purposes of this compendium. In terms of the skill areas that are the focus of this guide, the scale maps onto Communication. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

Ratings of OST program staff and teachers on the scale demonstrated satisfactory internal consistency reliability in samples of youth participating in OST programs in Massachusetts and Pennsylvania (staff ratings only). Estimates of internal reliability for male and female youth, for elementary, middle and high school youth, and for Black, Hispanic and White youth in these samples were satisfactory. Data on interrater or test-retest reliability are not available for either type of rater.

### Validity

In support of convergent validity, in analyses of data for a sample of 359 youth in grades 4 through 8 from 17 CCLC program sites across five Massachusetts school districts a substantial positive correlation was found between OST program staff and teacher ratings on the scale ( $r = .47$ ). Correlations of ratings on the scale for each type of informant with ratings of the other informant on the teacher and staff versions of the SAYO (Problem Solving, Engagement, Initiative, Behavior, Adult Relations) also were reported. For teacher ratings four of these cross-construct correlations were lower than the convergent validity correlation reported above. The margin of difference was small, however, and for OST program staff ratings only two of the correlations met this same criterion. Overall, findings thus did not provide robust support for the scale's discriminant validity. In further findings for this sample, staff ratings on the scale exhibited significant positive associations with three of five scales on the SAYO-Y that ask youth about their experiences in OST programs; ratings of teachers, but not staff, on the scale were associated significantly with SAYO-Y scales that ask youth to retrospectively report academic and personal/social gains associated with their program participation. In support of criterion validity, ratings of OST program staff on the scale also exhibited significant correlations in the expected positive direction with teacher ratings of the quality of the youth's work on homework assignments and academic performance on the SAYO-T. In path modeling analyses with a sample of 442 youth from the same study, teacher ratings of the scale were intervening variables in several statistically significant and theoretically interpretable indirect pathways that linked youth reports of their program experiences with teacher ratings of their academic performance on the SAYO-T.

In other research with youth participating in 78 OST programs in Massachusetts, a measure of program attendance was not found to be related significantly to change in teacher ratings on the scale over the course of a school year. In addition, change in teacher ratings on the scale did not vary significantly across program sites, despite evidence that the sites varied significantly in terms of their quality. In two other samples teacher and staff ratings on the measure, exhibited expected improvements over the course of a year in which a student was participating in an OST program.

## Summary

This scale shows evidence of being reliable when used by teachers or OST program staff to rate demographically varied groups of youth. The finding of a substantial association between teacher and program staff ratings of youth provides encouraging evidence of convergent validity. Available evidence does not, however, clearly establish the scale as a measure of communication skills distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of communication skills relative to measures that target skills in other areas. The scale has demonstrated concurrent associations with indices of youth school work and academic competence. It would be useful to build on these findings by examining the scale's association with outcomes in other areas, as well as those assessed at later points in youths' schooling or development (predictive validity). There is also evidence of expected associations with OST program participation for both teacher and staff ratings on the scale, although findings are mixed. Further investigation of the scale's sensitivity to effects of program participation using a quasi-experimental or randomized control design would be valuable. All available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

## Future Planning-My Actions

The Future Planning-My Actions scale of the SAYO-Y is a four-item self-report measure. The same items are used in the versions of the SAYO-Y that are intended for grades 4 through 8 and grades 9 through 12, respectively. In terms of the skill areas that are the focus of this guide, the content of this scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

The scale has demonstrated satisfactory internal consistency reliability within a large sample of youth in grades 6 through 12 participating in CCLC program sites in Massachusetts and a sample of youth in grades 9 and above from seven Philadelphia area youth programs. Internal consistency estimates reported within both samples for male and female youth and in the Philadelphia sample for youth in grades 6 through 8 and 9 through 12 and for White, Black and Hispanic youth also were satisfactory. No data are available on test-retest reliability.

### Validity

In analyses of data for 271 youth from 17 CCLC program sites in Massachusetts, the scale exhibited a significant positive association with OST program staff ratings on the Initiative scale of the SAYO-S ( $r = .37$ ). As the Future Planning-My Actions scale asks youth about their initiative and future-orientation in relation to school work, this finding provides support for the scale's convergent validity. However, the scale failed to exhibit an association with teacher ratings on the Initiative scale of the SAYO-T. In support of criterion validity, the scale exhibited expected positive associations with teacher ratings of the quality of the youth's work on homework assignments on the SAYO-T ( $r$ s of .24 and .21, respectively) and with youth reports on the SAYO-Y of their sense of competence ( $r$ s ranging from .33 to .41). For the same sample, the scale exhibited significant associations in expected directions with SAYO-Y scales that ask youth about their experiences in OST programs, as well as those that assess retrospective reports of academic and personal/social gains associated with program participation. In path modeling analyses conducted for a sample of 442 youth from the same study, ratings on the scale were similarly found to be intervening variables in several statistically significant and theoretically interpretable indirect pathways that linked youth reports of their program experiences with teacher ratings of youths' academic performance on the SAYO-T.

For the sample of youth in grades 6 through 12 referred to under Reliability, change in scores from fall to spring of a school year did not vary significantly across 36 OST program sites, despite evidence of differences in program quality in the form of significant variability in SAYO-Y scales assessing youth program experiences in various areas (e.g., challenge and engagement). Scores on the scale also did not change significantly from fall to spring assessments, thus failing to reflect evidence of sensitivity to effects of OST program participation. Youth reports of program experiences did, however, concurrently predict scores on the scale in expected directions. Finally, for the sample of youth participating in Philadelphia area programs referred to under Reliability, scores on the scale did not exhibit expected improvement between fall and spring assessments.

## Summary

This scale shows evidence of being reliable for demographically varied groups of youth. With regard to validity, available findings do not clearly establish the scale as a measure of future planning distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of the same construct (convergent validity) relative to measures that target skills or behaviors in other areas (discriminant validity). The scale has demonstrated concurrent associations with indices of youth school work and academic competence. It would be useful to build on these findings by examining the scale's associations with outcomes in other areas, as well as those assessed at later points in youths' schooling or development (predictive validity). The scale also has exhibited expected improvements in association with OST program participation and preliminary evidence of being a mediator of effects of program experiences on youth outcomes. To add to findings in this area, investigation of the scale's sensitivity to effects of OST program participation using an experimental or quasi-experimental design would be valuable. Currently, all available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

## Sense of Competence Socially

The Sense of Competence Socially scale of the SAYO-Y is a four-item self-report measure. There are separate versions of the scale for youth in grades 4 through 8 and grades 9 through 12, but the only difference is the substitution of “teens” for “kids” in the older version. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale’s reliability is substantial and evidence for the scale’s validity is moderate to substantial.

### Reliability

The Sense of Competence Socially scale of the SAYO-Y is a four-item self-report measure. There are separate versions of the scale for youth in grades 4 through 8 and grades 9 through 12, but the only difference is the substitution of “teens” for “kids” in the older version. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale’s reliability is substantial and evidence for the scale’s validity is moderate to substantial.

### Validity

In support of convergent validity, in a sample of 271 youth from 17 CCLC program sites in Massachusetts the scale exhibited a positive and significant association with program staff ratings on the Relations with Peers scale of the SAYO-S ( $r = .22$ ). The scale exhibited correlations ranging from .16 to .26 with other scales of the SAYO as rated by OST program staff (Behavior, Initiative, Engagement, Learning Skills, Communication Skills) and by teachers (Behavior, Initiative, Engagement, Problem Solving, Communication Skills, Homework). These associations are similar in magnitude to the convergent validity correlation noted above and thus do not provide clear support for the scale’s discriminant validity.

For the same sample, the scale exhibited significant associations in expected directions with SAYO-Y scales that ask youth about their experiences in OST programs, including a theoretically predictable relatively strong association with their perceptions of the program’s social environment ( $r = .57$ ). It also was associated in expected directions with SAYO-Y scales that ask youth for their retrospective reports of academic and personal/social gains associated with program participation. As would be predicted, the scale’s association with ratings of gains in the latter, socially oriented domain was relatively stronger ( $r$ s of .37 and .52, respectively). For the sample of youth in grades 4 through 8 referred to above under Reliability, youth reports of program experiences on the SAYO-Y again were concurrently related to scores on the scale in expected directions. Change in scores from fall to spring, however, did not vary significantly across 43 program sites, despite evidence of relevant differences in program quality in the form of significant variability in SAYO-Y scales assessing the program experiences of youth in various areas (e.g., social environment). Average scores on the scale also decreased significantly from fall to spring, which is unexpected, although the magnitude of the change was small. Finally, in path modeling analyses with a sample of 442 youth from the same study, the scale was an intervening variable in several statistically significant and theoretically interpretable pathways that linked youth reports of their OST program experiences on the SAYO-Y to teacher ratings of their academic competence on the SAYO-T.

### Summary

This scale shows evidence of being reliable for demographically varied groups of youth. The finding of an association with program staff ratings of youth’s peer relations, although limited in strength, provides encouraging evidence of convergent validity. Available evidence does not, however, clearly establish the scale as a measure of social competence distinct from other constructs. To address this concern, it would be useful to investigate the scale’s associations with well-validated measures of social competence relative to measures that target skills or behaviors in other areas. Evidence of criterion validity is currently limited to associations with academic outcomes. It would also be useful in future research to examine the scale’s associations with youth outcomes in other relevant areas, such as mental health or workplace success. There is also preliminary, but mixed, evidence of expected associations with OST program participation for the scale. To add to findings in this area, investigation of the scale’s sensitivity to effects of program participation using a quasi-experimental or randomized control design would be valuable. All available findings involve other scales within the SAYO assessment system. It would be helpful in future research on the scale to include well-validated measures from other sources.

# Technical Summaries of Scales Reviewed from the *Youth Experiences Survey 2.0 (YES, 2.0)*

## Goal Setting

The Goal Setting scale of the YES, 2.0 is a three-item self-report measure for middle- and high school-age youth. This scale – along with Effort, Problem Solving and Time Management – comprise the broader Initiative Experiences scale of the YES. It corresponds to the Initiative and Self-Direction focus area for this guide. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Initiative Experiences, is also reported. Evidence for reliability for this scale is moderate and evidence for the scale's validity is limited to moderate.

### Reliability

Ratings from youth in middle school and high school on this scale have demonstrated satisfactory internal consistency reliability in geographically diverse samples. Information about test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the broader Initiative Experiences scale exhibited substantial associations ( $r$ s of .52, .65, .71, .61, and .57) with all other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for Initiative Experiences, although not necessarily ratings of goal-setting experiences specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's experiences pertaining to goal setting were found to be significantly associated with the youth's own ratings on the goal-setting scale ( $r = .41$ ). In support of criterion validity, in a study of 1,504 Australian youth, two-level hierarchical linear models (students nested within schools) found that scores on the goal-setting scale with respect to youths' experiences in sports were predictive of more favorable youth reports of general self-worth. This association was stronger for youth from low-SES schools. Another study utilizing path analyses with a sample of 119 youth ages 10 through 18 found indirect pathways from youth reports of perceived competence at baseline and their reports of goal setting experiences six weeks later.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to goal-setting distinct from other areas of program experience, especially those that comprise the remainder of the broader Initiative Experiences scale of the YES. Evidence for convergent validity is promising, but limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of goal-setting in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize goal-setting.

## Problem-Solving

The Problem Solving scale of the YES, 2.0 is a three-item self-report measure for middle- and high school-age youth. This scale – *along with* Effort, Goal Setting, and Time Management – comprise the broader Initiative Experiences scale of the YES. It corresponds to the Critical Thinking and Decision-Making focus area for this guide. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Initiative Experiences, is also reported. Evidence for reliability for this scale is not available and evidence for the scale's validity is limited.

### Reliability

Information about internal consistency and test-retest reliability for the scale is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the broader Initiative Experiences scale exhibited substantial associations (rs of .52, .65, .71, .61, and .57) with all other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for Initiative Experiences although not necessarily ratings of problem-solving experiences specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's experience of problem solving were found to be significantly, although not strongly, associated with the youth's own ratings on the problem-solving scale ( $r = .27$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to problem solving distinct from other areas of program experience, especially those that comprise the remainder of the broader Initiative Experiences scale of the YES. Evidence for convergent validity is somewhat promising, but limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of problem-solving in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize problem solving.



## Time Management

The Time Management scale of the YES, 2.0 is a three-item self-report measure for middle- and high school-age youth. This scale – along with Effort, Problem Solving and Goal Setting – comprise the broader Initiative Experiences scale of the YES. It corresponds to the Initiative and Self-Direction focus area for this guide. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Initiative Experiences, is also reported. Evidence for reliability for this scale is not available and evidence for the scale's validity is limited.

### Reliability

Information about internal consistency and test-retest reliability for the scale is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the broader Initiative Experiences scale exhibited substantial associations ( $r$ s of .52, .65, .71, .61, and .57) with all other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for Initiative Experiences although not necessarily ratings of time management experiences specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's experiences pertaining to problem solving were found to be significantly associated with the youth's own rating ( $r = .35$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to time management distinct from other areas of program experience, especially those that comprise the remainder of the broader Initiative Experiences scale of the YES. Evidence for convergent validity is somewhat promising but is limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of time management in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize time management.

## Effort

The Effort scale of the YES, 2.0 is a three-item self-report measure for middle- and high school-age youth. It, along with Time Management, Problem Solving and Goal Setting, comprise the broader Initiative Experiences scale of the YES. It corresponds to the Initiative and Self-Direction focus area for this guide. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Initiative Experiences, is also reported. Evidence for reliability for this scale is not available and evidence for the scale's validity is limited.

### Reliability

Information about internal consistency and test-retest reliability for the scale is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the broader Initiative Experiences scale exhibited substantial associations (rs of .52, .65, .71, .61, and .57) with all other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for Initiative Experiences although not necessarily ratings of experiences of exerting effort specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 different programs, adult leaders' ratings of their perception of each youth's experience of effort were found to be significantly and fairly strongly associated with the youth's own rating ( $r = .51$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to time management distinct from other areas of program experience, especially those that comprise the remainder of the broader Initiative Experiences scale of the YES. Evidence for convergent validity is promising but is limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of effort in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize demonstrating effort.

## Emotion Regulation

The Emotion Regulation scale of the YES, 2.0 is a four-item self-report measure for middle- and high school-age youth. It corresponds to the Initiative and Self-Direction focus area for this guide. Evidence for reliability for this scale is limited and evidence for the scale's validity is limited.

### Reliability

Ratings from youth in high school on this scale have demonstrated satisfactory internal consistency reliability. Information about test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth scores on the scale were not associated with three negative experiences measured by the YES (inappropriate adult behavior, negative influence and social exclusion) and significantly, but only slightly, correlated with negative group dynamics and stress. The scale exhibited substantial associations ( $r$ s of .52- .65) with all other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity. Convergent validity was not demonstrated when, across 16 different programs, adult leaders' ratings of their perception of each youth's experiences pertaining to emotion regulation were not found to be associated significantly with the youth's own rating on the emotion regulation scale.

### Summary

Evidence for the scale's reliability is encouraging but is limited by a lack of information on reliability for ratings of demographic subgroups of youth (age, gender, racial/ethnic, socioeconomic). With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to emotion regulation distinct from other areas of program experience. With regard to convergent validity, a single study did not find any association between youth self-perceptions of improvements in emotional regulation and adult leaders' observations or perceptions of improvements in emotional regulation. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of emotional regulation in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that intentionally address emotion regulation.

## Diverse Peer Relationships

The Diverse Peer Relationships scale of the YES, 2.0 is a four-item self-report measure for middle- and high school-age youth. It, along with the Prosocial Norms scale, comprise the broader Positive Relationships scale of the YES. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Positive Relationships, is also reported. Evidence for reliability for this scale is not available and the evidence for the scale's validity is limited.

### Reliability

Information about the scale's internal consistency and test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the broader Positive Relationships scale exhibited substantial associations ( $r$ s of .54 - .63) with the other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for Positive Relationships although not necessarily ratings of experiences of Diverse Peer Relationships specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's experience of diverse peer relationships were found to be significantly, although not strongly, associated with the youth's own rating ( $r = .22$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to diverse peer relationships distinct from other areas of program experience, especially those that comprise the remainder of the broader Positive Relationships scale of the YES. Findings from a single study provide limited support for convergent validity. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of cultural competence or sensitivity in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize diverse peer relationships.

## Prosocial Norms

The Prosocial Norms scale is a four-item self-report measure for middle- and high school-age youth. It, along with the Diverse Peer Relationships scale, comprise the broader Positive Relationships scale of the YES. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Positive Relationships, is also reported. Evidence for reliability for this scale is not available and evidence for the scale's validity is limited to moderate.

### Reliability

Information about the scale's internal consistency and test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the broader Positive Relationships scale exhibited substantial associations ( $r$ s of .54- .63) with the other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for Positive Relationships although not necessarily ratings of experiences of Prosocial Norms specifically. In support of convergent validity, in a sample of 118 seventh-through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's experience of diverse peer relationships were found to be significantly associated with the youth's own rating ( $r = .40$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to prosocial norms distinct from other areas of program experience, especially those that comprise the remainder of the broader Positive Relationships scale of the YES. Evidence for convergent validity is promising, but limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of prosocial behavior in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize prosocial norms.

## Group Process Skills

The Group Process Skills scale of the YES, 2.0 is a five-item self-report scale for middle- and high school-age youth. It, along with Feedback and Leadership and Responsibility scales, comprise the broader Teamwork and Social Skills scale of the YES. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Teamwork and Social Skills, is also reported. Evidence of the scale's reliability is not available, and evidence for the scale's validity is limited to moderate.

### Reliability

Information about the scale's internal consistency and test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the Teamwork and Social Skills scale exhibited substantial associations ( $r$ s of .55 to .71) with all other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for the broader Teamwork and Social Skills scale although not necessarily ratings of group process skills specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's experience of group process skills were found to be significantly associated with the youth's own rating ( $r = .49$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to group process skills distinct from other areas of program experience, especially those that comprise the remainder of the broader Teamwork and Social Skills scale of the YES. Evidence for convergent validity is promising, but limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of group process skills in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize group process skills.

## Feedback

The Feedback scale of the YES is a two-item self-report scale for middle- and high school-age youth. It, along with Group Process Skills and Leadership and Responsibility, comprise the broader Teamwork and Social Skills scale of the YES. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Teamwork and Social Skills, is also reported. Evidence of the scale's reliability is not available and evidence for the scale's validity is limited to moderate.

### Reliability

Information about the scale's internal consistency and test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the Teamwork and Social Skills scale exhibited substantial associations ( $r$ s of .55, .to .71) with the other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for the broader Teamwork and Social Skills scale although not necessarily ratings of feedback experiences specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 programs, adult leaders' ratings of their perception of each youth's feedback experiences were found to be significantly associated with the youth's own rating ( $r = .35$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to feedback distinct from other areas of program experience, especially those that comprise the remainder of the broader Teamwork and Social Skills scale of the YES. Evidence for convergent validity is somewhat promising, but limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of skills for giving and receiving feedback in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize using or providing feedback.

## Leadership and Responsibility

The Leadership and Responsibility scale of the YES, 2.0 is a three-item self-report scale for middle- and high school-age youth. It, along with Feedback and Group Process Skills, comprise the broader Teamwork and Social Skills scale of the YES. As little analysis has been done with the YES at the lowest scale level, relevant validity information for the combined scale, Teamwork and Social Skills, is also reported. Evidence of the scale's reliability is not available and evidence for validity is limited to moderate.

### Reliability

Information about the scale's internal consistency and test-retest reliability is not available.

### Validity

With regard to discriminant validity, in a sample of 1,822 eleventh-grade youth, the Teamwork and Social Skills scale exhibited substantial associations ( $r$ s of .55, .to .71) with the other positive scales in the YES. However, confirmatory factor analysis indicated multi-factor scales fit the data better than a single positive dimension, which is consistent with some discriminant validity for the broader Teamwork and Social Skills scale although not necessarily ratings of leadership and responsibility experiences specifically. In support of convergent validity, in a sample of 118 seventh- through 12<sup>th</sup>-grade youth from 16 different programs, adult leaders' ratings of their perception of each youth's experience of leadership and responsibility were found to be significantly associated with the youth's own rating ( $r = .47$ ).

### Summary

Evidence for the scale's reliability is not available. With regard to validity, available findings do not clearly establish the scale as a measure of experiences relating to leadership and responsibility distinct from other areas of program experience, especially those that comprise the remainder of the broader Teamwork and Social Skills scale of the YES. Evidence for convergent validity is promising, but limited to a single study. To further address these concerns, it would be useful to investigate the scale in relation to well-validated measures of youth program experiences other than the YES. Also helpful would be longitudinal studies that test the relationship between scores on the scale and measures of leadership and responsibility in other settings at later points in youths' schooling or development (predictive validity). Examining change in scores on the scale over the course of program participation would be helpful as well, especially with respect to whether gains are differentially more evident when youth are participating in programs that emphasize leadership and responsibility.



# Technical Summaries of Scales Reviewed from the *Youth Outcomes Battery*

## Friendship Skills

The Friendship Skills scale of the *Camp Youth Outcome Scales - Detailed Version (CYOS-DV): Status Format* is a 13-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is limited as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a pilot study conducted with 391 youth attending four summer camps. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest or interrater reliability.

### Validity

In the pilot study referred to above, scores on the scale exhibited significant associations with the Teamwork and Independence status format scales of the *CYOS-DV* ( $r_s = .67$  and  $.52$ , respectively). The noteworthy magnitude of these associations is not suggestive of a high level of discriminant validity for the scale. As would be expected, however, scores on the scale were associated significantly with youth ratings of their change since coming to camp on the same items ( $r = .49$ ); this association, furthermore, was stronger than the scale's corresponding associations with ratings of change in the areas of teamwork and independence ( $r_s = .31$  and  $.36$ , respectively).

### Summary

Support for the scale's reliability is encouraging but is limited to a single sample and does not include information on reliability for different demographic subgroups of youth. Evidence for validity is similarly limited and restricted to associations with other self-report scales from the same instrument. It would be useful in future research to gather information regarding both the scale's association with well-validated measures of social competence (convergent validity) and its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It would also be helpful to investigate the scale's ability to predict relevant criterion measures, as well as its sensitivity to effects of OST program participation.

## Independence

The Independence scale of the *CYOS-DV: Status Format* is an eight-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is limited as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a pilot study conducted with 391 youth attending four different summer camps. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest or interrater reliability.

### Validity

In the pilot study referred to above, scores on the scale exhibited significant associations with the Teamwork and Friendship Skills status format scales of the *CYOS-DV* ( $r_s = .61$  and  $.52$ ). The noteworthy magnitude of these associations is not suggestive of a high level of discriminant validity for the scale. As would be expected, however, scores on the scale were associated significantly with youth ratings of their change since coming to camp on the same items ( $r = .37$ ); this association, furthermore, was marginally stronger than the scale's corresponding associations with ratings of change in the areas of teamwork and friendship skills ( $r_s = .31$ ).

## Summary

Support for the scale's reliability is encouraging but is limited to a single sample and does not include information on reliability for different demographic subgroups of youth. Evidence for validity is similarly limited and restricted to associations with other self-report scales from the same instrument. It would be useful in future research to gather information regarding both the scale's association with well-validated measures of independence (convergent validity) and its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It would also be helpful to investigate the scale's ability to predict relevant criterion measures, as well as its sensitivity to effects of OST program participation.

## Teamwork

The Teamwork scale of the *CYOS-DV: Status Format* is an eight-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is limited as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a pilot study conducted with 391 youth attending four different summer camps. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest or interrater reliability.

### Validity

In the pilot study referred to above, scores on the scale exhibited significant associations with the Independence and Friendship Skills status format scales of the *CYOS-DV* ( $r_s = .61$  and  $.67$ , respectively). The noteworthy magnitude of these associations is not suggestive of a high level of discriminant validity for the scale. As would be expected, however, scores on the scale were associated significantly with youth ratings of their change since coming to camp on the same items ( $r = .48$ ); this association, furthermore, was stronger than the scale's corresponding associations with ratings of change in the areas of independence and friendship skills ( $r_s = .27$  and  $.34$ , respectively).

### Summary

Support for the scale's reliability is encouraging but is limited to a single sample and does not include information on reliability for different demographic subgroups of youth. Evidence for validity is similarly limited and is restricted to associations with other self-report scales from the same instrument. It would be useful in future research to gather information regarding both the scale's association with well-validated measures of teamwork or cooperation (convergent validity) and its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It would also be helpful to investigate the scale's ability to predict relevant criterion measures, as well as its sensitivity to effects of OST program participation.

## Interest in Exploration

The Interest in Exploration scale of the *CYOS-DV: Status Format* is an eight-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is limited as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a pilot study conducted with 92 youth attending three different summer camps. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest or interrater reliability.

### Validity

In the pilot study referred to above, scores on the scale exhibited significant associations with the Perceived Competence, Responsibility and Family Citizenship Behavior status format scales of the *CYOS-DV* ( $r_s = .74$ ,  $.51$ , and  $.53$ , respectively). The noteworthy magnitude of these associations is not suggestive of a high level of discriminant validity for the scale. As would be expected, however, scores on the scale were associated significantly with youth ratings of their change since coming to camp on the same items ( $r = .40$ ); this association, furthermore, was stronger than the scale's corresponding associations with ratings of change in the other areas noted above ( $r_s$  ranging from  $.13$  to  $.17$ ).

### Summary

Initial support for the scale's reliability is encouraging. Findings are limited to a single relatively small sample, however, and do not include information on reliability for different demographic subgroups of youth. Validity evidence is similarly limited and is restricted to associations with other self-report scales from the same instrument. It would be useful in future research to gather information regarding both the scale's association with well-validated measures of comparable constructs (convergent validity) and its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It would also be helpful to investigate the scale's ability to predict relevant criterion measures, as well as its sensitivity to effects of OST program participation.

## Responsibility

The Responsibility scale of the *CYOS-DV: Status Format* is a six-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is limited and evidence for the scale's validity is none-to-limited.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a pilot study conducted with 92 youth attending three different summer camps. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest or interrater reliability.

### Validity

In the pilot study referred to above, scores on the scale exhibited significant associations with the Perceived Competence, Interest in Exploration and Family Citizenship Behavior status format scales of the *CYOS-DV* ( $r_s = .53, .51, \text{ and } .39$ , respectively). The fairly substantial magnitude of these associations is not suggestive of a high level of discriminant validity for the scale. Furthermore, whereas scores on the scale generally would be expected to exhibit a significant association with youth ratings of their change since coming to camp on the same items, this was not the case ( $r = .04$ ).

### Summary

Initial support for the scale's reliability is encouraging. Findings are limited to a single, relatively small sample, however, and do not include information on reliability for different demographic subgroups of youth. Validity evidence is similarly limited and is restricted to associations with other self-report scales from the same instrument. It would be useful in future research to gather information regarding both the scale's association with well-validated measures of comparable constructs (convergent validity) and its degree of overlap with measures of conceptually distinct constructs (discriminant validity). It would also be helpful to investigate the scale's ability to predict relevant criterion measures, as well as its sensitivity to effects of OST program participation.

## Problem-Solving Confidence

The Problem-Solving Confidence scale of the *CYOS-DV: Status Format* is an eight-item self-report measure. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Critical Thinking and Decision-Making. Evidence for the scale's reliability is limited as is evidence for the scale's validity.

### Reliability

This scale has demonstrated satisfactory internal consistency reliability in a pilot study conducted with 221 youth attending residential summer camps. Internal consistency reliability estimates for demographic subgroups of youth are not available nor is information on test-retest or interrater reliability.

### Validity

In support of the scale's convergent validity, in the pilot study referred to above, scores on the scale exhibited a significant correlation ( $r = .60$ ) with the Problem-Solving Skills Subscale (PSSS) of the *Social Problem Solving Inventory for Adolescents*. As also would be expected, scores on the scale were associated significantly with youth ratings of their change since coming to camp on the same items ( $r = .59$ ).

### Summary

Initial support for the scale's reliability is encouraging. Findings are limited to a single sample, however, and do not include information on reliability for different demographic subgroups of youth. Validity evidence is similarly encouraging but also limited. It would be useful in future research to gather further information regarding the scale's association with well-validated indices of problem-solving ability including, ideally, measures that are based on observer ratings or objective assessment. It also would be helpful to examine the scale's differentiation from measures of other constructs (discriminant validity), its ability to predict relevant youth outcomes (criterion validity), and its sensitivity to effects of OST program participation.

# Technical Summaries of Scales Reviewed from the *Youth Outcome Measures Online Toolbox*

## Social Skills

The Social Skills scale of the *Youth Outcome Measures Online Toolbox* is a seven-item measure completed by the youth's teacher or OST program staff. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate.

### Reliability

Both teacher and OST program staff ratings on this scale have demonstrated satisfactory internal consistency reliability in geographically diverse samples of elementary school and middle school youth. In a study of a large sample of elementary school (grades 3 through 5) and middle school (grades 6 through 8) youth participating in OST programs in geographically varied locations, the scale's internal consistency reliability was satisfactory for both types of raters across three periods of assessment for males and females; elementary and middle school youth; low-income and non-low-income youth; English Language Learner (ELL) and non-ELL youth; and White, Black, Hispanic and Asian-American youth. Data on test-retest reliability are not available.

### Validity

In support of convergent validity, for the sample described above moderate correlations were found between OST program staff and teacher ratings on the scale across three periods of assessment for both elementary school youth (mean  $r = .38$ ) and middle school youth (mean  $r = .44$ ). Associations of teacher and OST program staff ratings on the scale with ratings of the same informant on the Prosocial Behavior scale that is also part of the *Online Toolbox* were moderate to strong ( $r_s = .63$  to  $.77$ ) and thus do not provide robust support for the scale's discriminant validity relative to other related constructs. Cross-informant associations between scores on the two scales were moderate ( $r_s = .27$  to  $.54$ ). In support of criterion validity, ratings on the scale were associated negatively with *Online Toolbox* ratings by teachers and staff of aggressive behavior. Of particular note are cross-informant associations in which teacher ratings on the scale were, as would be expected, related inversely to staff ratings of aggressive behavior ( $r_s = -.33$  to  $-.38$  and  $-.27$  to  $-.43$  for elementary and middle school students, respectively) and vice-versa ( $r_s = -.32$  to  $-.33$  and  $-.35$  to  $-.47$  for elementary and middle school students, respectively). In the same study, change in teacher and OST program staff ratings on the scale were compared over one- and two-year periods (results for staff ratings are reported only for the one-year analyses) for cluster analysis-derived groups of students who had high levels of participation in OST programs judged to be high quality either with or without high levels of involvement in other organized OST activities, relative to students who experienced high levels of time after school in unsupervised activity. For elementary school youth, for the one-year analyses, teacher ratings on the scale improved to a relatively greater extent for youth in the high program/low activity cluster and OST program staff ratings improved to a relatively greater extent for this cluster and for the high program/high activity cluster, whereas neither comparison (either high program/high activity or high program/low activity vs. high unsupervised) was significant for middle school youth for teacher or staff ratings. A similar pattern of findings was obtained in the two-year analyses.

In a longitudinal study of 150 youth participating in OST programs, an observation-based measure of cumulative program quality across grades 1 through 4 was associated with relatively greater improvement in teacher ratings on the scale over the same time period, controlling for measures of family income and parenting, among girls, but not among boys. In this study, teacher ratings on the scale of children in grades 1 through 4 also were associated inversely, as would be expected, with time sample observations of the children's negative peer interactions in OST programs ( $r_s = -.23$  to  $-.40$ ).

## Summary

There is strong evidence for the scale's reliability across a range of important demographic groups. Available findings do not clearly establish the scale as a measure of social skills distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of social skills (convergent validity) relative to measures that target abilities or skills in other areas (discriminant validity). In view of the noteworthy strength of the scale's associations with ratings of the same informant on the Prosocial Behavior scale of the *Online Toolbox*, for example, whether the scale is sufficiently specific to be regarded as a measure of a social skills or abilities as distinct from a youth's observed tendencies to actually engage in positive behavior with peers would benefit from further examination. The scale has exhibited expected associations with indices of aggressive and other negative behavior (concurrent validity). It would be useful to add to these findings by examining the scale's ability to contribute to prediction of similar outcomes for youth at later points in their schooling or development (predictive validity). Evidence of the scale's sensitivity to effects of OST program participation is encouraging. In future research, it would be helpful to further examine this issue among older, middle school- and high school-age samples and to incorporate the use of a randomized control evaluation design.



## Prosocial Behavior

The Prosocial Behavior scale of the *Youth Outcome Measures Online Toolbox* is an eight-item measure completed by the youth's teacher or OST program staff. The items comprise those of the subscale by the same name in the *Child Behavior Scale (CBC)* (Ladd & Profilet, 1996); the scale also makes use of the *CBC*'s three-point response scale, although with somewhat different descriptors for response options. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is substantial and evidence for the scale's validity is moderate to substantial.

### Reliability

Both teacher and OST program staff ratings on this scale have demonstrated satisfactory internal consistency reliability in geographically diverse samples of elementary school and middle school youth. In a study of a large sample of elementary school (grades 3 through 5) and middle school (grades 6 through 8) youth participating in OST programs in geographically varied locations, the scale's internal consistency reliability was satisfactory for both types of raters across three periods of assessment for males and females; elementary and middle school youth; low-income and non-low-income youth; English Language Learner (ELL) and non-ELL youth; and White, Black, Hispanic and Asian-American youth. Data on test-retest reliability are not available.

### Validity

In support of convergent validity, for the sample described above moderate correlations were found between OST program staff and teacher ratings on the scale across three periods of assessment for both elementary school youth (mean  $r = .34$ ) and middle school youth (mean  $r = .41$ ). Associations of teacher and OST program staff ratings on the scale with ratings of the same informant on the Prosocial Behavior scale that is also part of the *Online Toolbox* were moderate to strong ( $r_s = .63$  to  $.77$ ) and thus do not provide robust support for the scale's discriminant validity relative to other, related constructs. In the same study, change in teacher and OST program staff ratings on the scale were compared over one- and two-year periods (results for staff ratings are reported for only the one-year analyses) for cluster analysis-derived groups of students who had high levels of participation in OST programs judged to be high quality either with or without high levels of involvement in other organized OST activities, relative to students who experienced high levels of time after school in unsupervised activity. For elementary school youth, for the one-year analyses, teacher ratings on the scale improved to a relatively greater extent for youth in the high program/low activity cluster and OST staff ratings improved to a relatively greater extent for this cluster and for the high program/high activity cluster, whereas neither comparison (either high program/high activity or high program/low activity vs. high unsupervised) was significant for middle school youth for either teacher or staff ratings. A similar pattern of findings was obtained for the two-year analyses.

In a study that followed 391 children from grades 1 through 8, in support of convergent validity, teacher ratings on the subscale of the *CBC* that serves as the basis for the Prosocial Scale exhibited significant and substantial associations with a peer nomination measure of prosocial behavior at each of three time points (grade 4:  $r = .37$ ; grade 6:  $r = .52$ , grade 8:  $r = .54$ ). In support of discriminant validity, teacher ratings on the scale exhibited a pattern of relatively weaker associations with both teacher and peer nominations of other facets of peer relations (asocial with peers, excluded by peers). In support of criterion validity, in this same study a latent variable used to represent teacher ratings on the scale at grade 5 was predictive of a latent measure of relational competence derived from peer rating-based measures (e.g., peer group acceptance) at both grade 6 and grade 8. In addition, teacher ratings on the scale were correlated concurrently in the expected negative direction with parent- and teacher-report measures of aggressive behavior at grades 4, 6, and 8 ( $r_s$  ranging from  $-.29$  to  $-.57$ ).

## Summary

There is strong evidence for the scale's reliability across a range of important demographic groups. Evidence of the scale's convergent and discriminant validity is generally encouraging. In view of the scale's noteworthy associations with the Social Skills scale on the same instrument, however, whether the scale is sufficiently specific to be regarded as a measure of a youth's observed tendencies to actually engage in positive behavior with peers as distinct from the youth's social skills or abilities would benefit from additional examination. The scale's criterion validity has received support via expected associations with indices of youth social functioning and behavior both concurrently and over time. Evidence of the scale's sensitivity to effects of OST program participation is also encouraging. In future research, it would be helpful to further examine this issue among older, middle school- and high school-age samples and to incorporate the use of a randomized control evaluation design.

## Work Habits

The Work Habits scale of the *Youth Outcome Measures Online Toolbox* is a six-item measure that is completed by either the youth's teacher or OST program staff or by the youth. For purposes of this guide, technical properties of the adult- and youth-report versions of the scale were evaluated separately. In terms of the skill areas that are the focus of this guide, the contents of the scale maps onto Initiative and Self-Direction.

### Teacher- and OST Program Staff-Report Version

The teacher and OST program staff versions of the scale are intended to assess the youth's work habits in the classroom or an OST program, respectively. For these versions of the scale, evidence to date for reliability is substantial and evidence for validity is limited to moderate.

### Reliability

Both teacher and OST program staff ratings on this scale have demonstrated satisfactory internal consistency reliability in geographically diverse samples of elementary school and middle school youth. In a study of a large sample of elementary school (grades 3 through 5) and middle school (grades 6 through 8) youth participating in OST programs in geographically varied locations, internal consistency reliability of the scale was satisfactory for both types of raters across three periods of assessment for males and females, elementary and middle school youth; low-income and non-low-income youth; English Language Learner (ELL) and non-ELL youth; and White, Black, Hispanic and Asian-American youth. Data on test-retest reliability are not available.

### Validity

In support of convergent validity, for the sample described above, moderate correlations were found between OST program staff and teacher ratings on the scale across three periods of assessment for both elementary school youth (mean  $r = .44$ ) and middle school youth (mean  $r = .50$ ). In this same study, change in teacher and OST program staff ratings on the scale were compared over one and two-year periods (results for staff ratings are reported for only the one-year analyses) for cluster analysis-derived groups of students who had high levels of participation in OST programs judged to be high quality either with or without high levels of involvement in other organized OST activities relative to students who experienced high levels of time after school in unsupervised activity. For elementary school youth, for the one-year analyses, teacher ratings on the scale improved to a relatively greater extent for youth in the high program low activity cluster and OST program staff ratings improved to a relatively greater extent for this cluster and for the high program/high activity cluster. At the middle school level, teacher ratings on the scale improved to a relatively greater extent for youth in the high program/high activity cluster, but there was no differential change across clusters for staff ratings. For two-year analyses, teacher ratings improved to a relatively greater extent for each program cluster at the elementary school level, but no differences were found at the middle school level. In the same study, teacher ratings on the scale exhibited substantial associations with teacher reports of the youth's academic performance at both the elementary school ( $r_s = .63$  to  $.68$ , over the three periods of assessment) and middle school ( $r_s = .59$  to  $.64$ ) levels. Smaller, but still noteworthy associations were found between OST program staff ratings on the scale and the same teacher measure of academic performance ( $r_s = .23$  to  $.28$  and  $.29$  to  $.55$  at the elementary and middle school levels, respectively).

In a longitudinal study of extracurricular activity participation among a sample of 1,364 children from 10 locations in the U.S. from kindergarten through fifth-grade, those with moderate- and high-intensity involvement exhibited significantly greater improvement in scores on the scale compared to those with low-intensity involvement. In a longitudinal study of 150 youth participating in OST programs, an observation-based measure of cumulative program quality across grades 1 through 4 was associated with relatively greater improvement in teacher ratings on the scale over the same time period, controlling for measures of family income and parenting, among girls, but not among boys. In this same study, teacher ratings on the scale exhibited a substantial association ( $r = .65$ ) with teacher ratings of academic performance at both Grade 4 and Grade 5.

## Summary

There is strong evidence for the scale's reliability across a range of important demographic groups. Available findings do not clearly establish the scale as a measure of work habits distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of work habits (convergent validity) relative to measures that target abilities or skills in other areas (discriminant validity). In view of the notably strong associations that teacher ratings on the scale have exhibited with teacher reports of academic performance, the extent to which the scale as completed by teachers is sufficiently specific to be regarded as a measure of a youth's work habits at school distinct from the youth's academic performance would benefit from additional investigation. Similarly, although the associations found with reports of academic performance provide support for the scale's concurrent validity, it would be helpful also to examine the scale's ability to predict improvements in academic achievement over time. Evidence of the scale's sensitivity to effects of OST program participation is encouraging. In future research, it would be helpful to further examine this issue among older, middle school- and high school-age samples and to incorporate the use of a randomized control evaluation design.

## Youth-Report Version

For the youth-report version of the scale, evidence to date for reliability is moderate to substantial and evidence for validity is limited to moderate.

## Reliability

This scale has demonstrated satisfactory internal consistency reliability in geographically diverse samples of elementary, middle- and high school-age youth. In a study of a large sample of elementary school (grades 3 through 5) and middle school (grades 6 through 8) youth participating in OST programs in geographically varied locations, the scale's internal consistency reliability across three periods of assessment was generally satisfactory for males and females; elementary and middle school youth, low-income and non-low-income youth; English Language Learner (ELL) and non-ELL youth; and White, Black, Hispanic and Asian-American youth. Test-retest reliability over a three- to four-month period for a sample of 238 youth in grades 2 through 8 was moderate ( $r = .54$ ).

## Validity

In the same study described above that provided subgroup estimates of reliability, change in youth ratings on the scale was examined over one and two-year periods for cluster analysis-derived groups of students who had high levels of participation in OST programs judged to be high quality either with or without high levels of involvement in other organized OST activities, relative to students who experienced high levels of time after school in unsupervised activity. For elementary school youth, for both the one and two-year analyses, ratings on the scale improved to a relatively greater extent for youth in both program clusters (high program/low activity and high program/high activity), whereas at the middle school level there were no significant differences. In this same study, youth ratings on the scale exhibited significant associations with teacher reports of academic performance at both the elementary school ( $r_s = .23$  to  $.29$ ) and middle school ( $r_s = .26$  to  $.32$ ) levels.

In a study of 246 students in grades 7 through 12 enrolled in an educationally focused OST program, student ratings on the scale were associated inversely, as would be expected, with number of days absent from the program. In the study of 238 youth in grades 2 through 8 referred to above, it was found that among males and middle school students, those with frequent attendance for the OST program being evaluated exhibited significantly greater relative improvement in their ratings on the scale than those in a low-attendance group.

## Summary

Overall, the evidence for the scale's reliability is strong and encompassing of a range of important demographic groups. Further investigation of test-reliability would be beneficial. Available findings do not clearly establish the scale as a measure of work habits distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of work habits (convergent validity) relative to measures that target abilities or skills in other areas (discriminant validity). Evidence of criterion-related validity is promising based on concurrent associations with teacher reports of academic performance. It would be useful to add to these findings by examining the scale's ability to predict similar outcomes for youth at later points in their schooling or development. Evidence of the scale's sensitivity to effects of OST program participation is encouraging. In future research, it would be helpful to further examine this issue among older, middle school- and high school-age samples and to incorporate the use of a randomized control evaluation design.

## Task Persistence

The Task Persistence scale of the *Youth Outcome Measures Online Toolbox* is an eight-item measure completed by the youth's teacher or OST program staff. The scale is adapted from a self-report measure of self-efficacy developed by Walker and Arbretton (2001). In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Initiative and Self-Direction. Evidence for the scale's reliability is substantial and evidence for the scale's validity is limited to moderate.

### Reliability

Both teacher and OST program staff ratings on the scale have demonstrated satisfactory internal consistency reliability in geographically diverse samples of elementary school and middle school youth. In a study of a large sample of elementary school (grades 3 through 5) and middle school (grades 6 through 8) youth participating in OST programs in geographically varied locations, the scale's internal consistency reliability was satisfactory for both types of raters across three periods of assessment for males and females; elementary and middle school youth; low-income and non-low-income youth; English Language Learner (ELL) and non-ELL youth; and White, Black, Hispanic and Asian-American youth. Data on test-retest reliability are not available.

### Validity

In support of convergent validity, for the sample described above moderate correlations were found between OST program staff and teacher ratings on the scale across three periods of assessment for both elementary school youth (mean  $r = .38$ ) and middle school youth (mean  $r = .51$ ). In this same study, change in teacher and OST program staff ratings on the scale were compared over one- and two-year periods (results for staff ratings are reported only for the one-year analyses) for cluster analysis-derived groups of students who had high levels of participation in OST programs judged to be high quality either with or without high levels of involvement in other organized OST activities, relative to students who experienced high levels of time after school in unsupervised activity. For elementary school youth, for the one-year analyses, teacher ratings on the scale improved to a relatively greater extent for youth in the high program/low activity cluster and OST program staff ratings improved to a relatively greater extent for both this cluster and the high program/high activity cluster; for middle school youth, neither comparison was significant (high program/high activity or high program/low activity vs. high unsupervised) for either teacher or staff ratings. A similar pattern of findings was obtained in the two-year analyses. In the same study, teacher ratings on the scale exhibited substantial associations with teacher reports of academic performance for students in both elementary school ( $r_s = .60$  to  $.64$ ) and middle school ( $r_s = .55$  to  $.63$ ). Smaller but still noteworthy associations were found between OST program staff ratings on the scale and the same teacher measure of academic performance ( $r_s = .24$  to  $.34$  at the elementary level,  $r_s = .30$  to  $.55$  at the middle school level).

### Summary

There is strong evidence for the scale's reliability across a range of important demographic groups. Available findings do not clearly establish the scale as a measure of social skills distinct from other constructs. To address this concern, it would be useful to investigate the scale's associations with well-validated measures of task persistence (convergent validity) relative to measures that target abilities or skills in other areas (discriminant validity). In view of the notably strong associations that teacher ratings on the scale have exhibited with teacher reports of academic performance, for example, the extent to which the scale as completed by teachers is sufficiently specific to be regarded as a measure of a youth's task persistence in the school context as distinct from the youth's academic performance would benefit from additional investigation. Similarly, although the associations found with reports of academic performance provide support for the scale's concurrent validity, it would be helpful also to examine the scale's ability to predict improvements in academic achievement over time. There is preliminary evidence of the scale's sensitivity to effects of OST program participation. In future research, it would be helpful to further examine this issue among older, middle school- and high school-age samples and to incorporate the use of a randomized control evaluation design.

## Social Competencies

The Social Competencies scale of the *Youth Outcome Measures Online Toolbox* is a seven-item self-report measure. The scale is adapted from the Social Self-Efficacy subscale on the *Self-Efficacy Questionnaire for Children (SEQ-C; Muris, 2001)*, which is intended to assess youths' perceived capability for peer relationships and assertiveness. The items parallel those of the original scale with minor rewording; the adapted scale has a four-point response scale rather than the five-point response scale that is used for the *SEQ-C*. In terms of the skill areas that are the focus of this guide, the content of the scale maps onto Relationships and Collaboration. Evidence for the scale's reliability is moderate to substantial and evidence for the scale's validity is moderate.

### Reliability

Both teacher and OST program staff ratings on the scale have demonstrated satisfactory internal consistency reliability in geographically diverse samples of elementary school, middle school and high school youth. In a field test conducted with students in grades 3 through 8 in different locations throughout the state of California, internal consistency reliability estimates were satisfactory for male and female youth and for samples with varying proportions of youth who were English Language Learners or from low-income families. Data on test-retest reliability are not available.

### Validity

In research conducted on the *SEQ-C* with four different samples of adolescents in the Netherlands and the U.S., items comprising the Social Self-Efficacy subscale of the instrument for the most part loaded on a distinct factor from other items on the measure that are intended to assess emotional and academic self-efficacy. In one of these samples, which consisted of 697 students from three middle schools and two high schools in a rural school district in the Southeastern U.S., higher scores on the scale correlated significantly and in expected directions with youth-report measures of internalizing symptoms ( $r = -.25$ ) and global life satisfaction ( $r = .35$ ), but not externalizing symptoms ( $r = -.05$ ). In a sample of 318 high school students from a different rural school district in the same region of the country, in support of convergent validity, a version of the original scale that omitted two items due to low or mixed factor loadings exhibited a significant association with the Friend Satisfaction subscale of the *Multidimensional Students' Life Satisfaction Scale* ( $r = .33$ ). Relative to this association, correlations with scales from the same instrument assessing satisfaction with living environment, family and school were somewhat weaker ( $r_s = .18$  to  $.27$ ), although the scale's association with the Self-Satisfaction scale was stronger in magnitude ( $r = .45$ ). For this same sample, the scale exhibited significant negative correlations with self-report measures of affective and anxiety problems ( $r_s$  of  $-.21$  and  $-.23$ , respectively), but not somatic, conduct, or oppositional defiant problems on the Youth Self Report of the *Child Behavior Checklist*; it also was not associated with grade-point average.

### Summary

Overall, the evidence for the scale's reliability is strong and encompassing of a range of important demographic groups. Available evidence for convergent, discriminant and criterion-related validity is generally promising. These findings, however, are based on different versions of the scale and in some instances non-U.S. samples. For these reasons, it would be helpful to have additional investigation of the scale in the form that is included in the *Youth Outcome Measures Online Toolbox*. Examination of the scale's ability to predict youth outcomes at later points in time (predictive validity) also would be helpful, as would investigation of its sensitivity to effects of OST program participation.

# Annotated Bibliography of Sources for Technical Appendix

## From Soft Skills to Hard Data: Measuring Youth Program Outcomes

### Resilience & Youth Development Module of the California Healthy Kids Survey

Hanson, T. L., & Kim, J. O. (2007). *Measuring resilience and youth development: the psychometric properties of the Healthy Kids Survey*. (Issues & Answers Report, REL 2007–No. 034). Washington, DC: U.S. Department of Education. Retrieved from [http://www.ies.ed.gov/ncee/edlabs/regions/west/pdf/REL\\_2007034.pdf](http://www.ies.ed.gov/ncee/edlabs/regions/west/pdf/REL_2007034.pdf).

### Developmental Assets Profile

Galligan, S. B., Barnett, R. V., Brennan, M. A., & Israel, G. D. (2010). Understanding the link between gender role conflict, resilience, and propensity for suicide in adolescent and emerging adult males. *International Journal of Men's Health*, 9(3), 201-210. doi:10.3149/jmh.0903.201.

Scales, P. C. (2011). Youth developmental assets in global perspective: Results from international adaptations of the developmental assets profile. *Child Indicators Research*, 4, 619–645. doi:10.1007/s12187-011-9112-8.

Strachan, L., Cote, J., & Deakin, J. (2009). An evaluation of personal and contextual factors in competitive youth sport. *Journal of Applied Sport Psychology*, 21, 340–355. doi:10.1080/10413200903018667.

Whiteman, S. D., Becerra-Bernard, J. M., & McHale, S. M. (2010). The nature and correlates of sibling influence in two-parent African American families. *Journal of Marriage and Family*, 72, 267–281. doi:10.1111/j.1741-3737.2010.00698.

Search Institute. (2013). *Developmental Assets Profile: User Manual*. Minneapolis, Minn.: Search Institute.

### Devereux Student Strengths Assessment

LeBuffe, P., Shapiro, V., & Naglieri, J. (2005). *Devereux student strengths assessment (DESSA) user's manual*. Lewisville, N.C.: KPress.

Nickerson, A. B., & Fishman, C. (2009). Convergent and divergent validity of the Devereux Student Strengths Assessment. *School Psychology Quarterly*, 24, 48–59. doi:10.1037/a0015147.



### **Holistic Student Assessment**

Noam, G., Malti, T., & Guhn, M. (2012). From clinical-developmental theory to assessment: The Holistic Student Assessment Tool. *International Journal of Conflict and Violence, Volume 6 (2)*, 201-213.

Robertson, D.L., Noam, G., & Blass, J. (2013, May). *The Holistic Student Assessment (HSA): identifying students who show above or below average socio-emotional risks and resiliencies*. A poster presented at the 21st Annual Conference of the Society for Prevention Research in San Francisco, CA.

Vandell, D.L., & Pierce, K.M. (2005). Study of After-School Care and Children's Development. Eunice Kennedy Shriver National Institute of Child Health and Human Development.

Vandell, D. L., Pierce, K. M., Brown, B., Lee, D., Bolt, D., Dadisman, K., Pechman, E., & Reisner, E. R. (2006, March). Developmental Outcomes Associated with the After-School Contexts of Low-Income Children and Youth: A study of Promising after-school programs and organized activities.

### **San Francisco Beacons Youth Survey**

Arbreton, A. J. A., Bradshaw, M., Metz, R., Sheldon, J., & Pepper, S. (2008). *More Time for Teens: Understanding Teen Participation—Frequency, Intensity and Duration—In Boys & Girls Clubs*. Philadelphia: Public/Private Ventures.

Arbreton, A. J. A., Bradshaw, M., Sheldon, J., & Pepper, S. (2009). *Making Every Day Count: Boys & Girls Clubs' Role in Promoting Positive Outcomes for Teens*. Philadelphia: Public/Private Ventures. Retrieved from [http://www.sp2.upenn.edu/ostrc/docs/document\\_library/cade/Adolescents%20and%20Youth%20Culture/Making%20Every%20Day%20Count%20-%20Teens.pdf](http://www.sp2.upenn.edu/ostrc/docs/document_library/cade/Adolescents%20and%20Youth%20Culture/Making%20Every%20Day%20Count%20-%20Teens.pdf).

Arbreton, A. J. A., & McClanahan, W. (2002). *Targeted Outreach: Boys and Girls Clubs of America's Approach to Gang Prevention and Intervention*. Philadelphia: Public/Private Ventures. Retrieved from: <http://www.goccp.maryland.gov/msac/documents/gang-studies/boys-and-girls-club/Arbreton-McClanahan-2005.pdf>.

Connell, J. P., Grossman, J. B., & Resch, N. (1995). *Urban Corps Assessment Package (URCAP) Manual*. Philadelphia: Public/Private Ventures. Retrieved from [http://www.issueLab.org/resource/urban\\_corps\\_assessment\\_package\\_urcap\\_manual](http://www.issueLab.org/resource/urban_corps_assessment_package_urcap_manual).

Walker, K. E., & Arbreton, A. J. A. (2004). *After-School Pursuits: An Examination of Outcomes in the San Francisco Beacon Initiative*. Philadelphia: Public/Private Ventures. Retrieved from [http://www.sfbeacon.org/00\\_Evaluations/Evaluations/1999-2001\\_After-School\\_Pursuits\\_Beacons\\_Outcome\\_Full\\_Report.pdf](http://www.sfbeacon.org/00_Evaluations/Evaluations/1999-2001_After-School_Pursuits_Beacons_Outcome_Full_Report.pdf).

### **Survey of Afterschool Youth Outcomes**

Miller, B. M., & Surr, W. (2003). *SAYO construct validity and internal consistency results*. National Institute on Out-of-School Time Programs.

Surr, W. B., Tracy, A. J., & Richer, A. M. (2010). *Psychometric testing results of the SAYO-Youth and SAYO-Staff survey based upon a teen population attending United Way of South Eastern Pennsylvania Programs*. National Institute on Out-of-School Time Programs.

### **History and Development of APAS Tools (related to the Survey of Afterschool Youth Outcomes)**

No author. (2005). *The Massachusetts After-School Research Study (MARS)*. Intercultural Center for Research in Education (INCRE) Arlington, Mass.

No author. (2010). *Exploring SAYO connections-Results from multi-district correlations*. National Institute on Out-of-School Time.

### **Social Skills Improvement System**

Gresham, F., & Elliott, S. N. (2007). *Social Skills Improvement System (SSIS) Rating Scales*. Thousand Oaks, Calif.: Pearson Publishing.

### **Youth Experiences Survey 2.0**

Coatsworth, J. D., & Conroy, D. E. (2009). The effects of autonomy-supportive coaching, need satisfaction, and self-perceptions on initiative and identity in youth swimmers. *Developmental Psychology*, 45(2), 320-328.

Hansen, D. M., & Larson, R. (2005). The youth experience survey 2.0: instrument revisions and validity testing. *Urbana-Champaign, Ill.: University of Illinois*.

Hansen, D. M., & Larson, R. W. (2007). Amplifiers of developmental and negative experiences in organized activities: Dosage, motivation, lead roles, and adult-youth ratios. *Journal of Applied Developmental Psychology*, 28(4), 360-374.

Larson, R., & Hansen, D. (2005). The Development of Strategic Thinking: Learning to Impact Human Systems in a Youth Activism Program. *Human Development*, 48(6), 327-349.

MacDonald, D. J., Côté, J., Eys, M., & Deakin, J. (2012). Psychometric properties of the Youth Experience Survey with young athletes. *Psychology of Sport and Exercise*, 13(3), 332-340.

Wang, C. K. J., Erickson, K., & Côté, J. (2012). Experience in competitive youth sport and needs satisfaction: the Singapore Story. *International Journal of Sport Psychology*, 43(1), 15-32.

Bloomfield, C. J., & Barber, B. L. (2011). Developmental experiences during extracurricular activities and Australian adolescents' self-concept: Particularly important for youth from disadvantaged schools. *Journal of Youth and Adolescence*, 40, 582-594.

### **Youth Outcomes Battery**

American Camp Association. (2011). *Camp Youth Outcomes Battery: Measuring Developmental Outcomes in Youth Programs* (2nd ed.). Martinsville, Ind.: American Camp Association.

Ellis, G., & Sibthorp, J. (2006). *Development and Validation of Age-Appropriate Measures of Camper Outcomes*. Report to the American Camp Association. University of Utah.

Sibthorp, J., & Browne, L. (2011). *Development and validation of the Spiritual Well-being Scale for use with the ACA Youth Outcomes Battery*. Technical report submitted for the American Camp Association.

### Youth Outcome Measures Online Toolbox

Ladd, G. W., Herald-Brown, S. L., & Andrews, R. K. (2009). The Child Behavior Scale (CBS) revisited: A longitudinal evaluation of CBS subscales with children, preadolescents, and adolescents. *Psychological Assessment, 21*(3), 325–339. doi: 10.1037/a0016205.

Lowe, D., Vandell, D. L., O'Cadiz, M. P., & Hall, V. (2009). *High quality supplemental educational services and afterschool partnerships demonstration project an evaluation study of think together programs in the Santa Ana unified school district*. University of California – Irvine.

Pierce, K. M., Lee, D., Vandell, D. L. (2007, March-April). Extracurricular participation and academic outcomes in the NICHD Study of Early Child Care and Youth Development. In L. M. Malone (Chair), *Learning outside of elementary school: A longitudinal look at extracurricular participation and achievement*. Paper symposium conducted at the biennial meetings of the Society for Research in Child Development, Boston.

Seldom, S. M., & Shaffer, E. J. (2007). Evaluation of the Self-Efficacy Questionnaire for Children in two samples of American adolescents. *Journal of Psychoeducational Assessment, 25*, 341–355. Doi: 10.1177/0734282907300636

Vandell, D. L. (2009). California Afterschool Outcome Measures Project: Phase I final report. Lucile Packard Foundation, Children, Families and Communities Program Grant # 2008 - 33187

Vandell, D. L., & Pierce, K. M. (2005). Study of After-School Care and Children's Development. Eunice Kennedy Shriver National Institute of Child Health and Human Development.

Vandell, D. L., & Pierce, K. M. (2001, April). Experiences in after-school programs and child well-being. In J. L. Mahoney (Chair), *Protective aspects of after-school activities: Processes and mechanisms*. Symposium conducted at the biennial meeting of the Society for Research in Child Development, Minneapolis, Minn.

Vandell, D. L., Pierce, K. M., Brown, B., Lee, D., Bolt, D., Dadisman, K., Pechman, E., & Reisner, E. R. (2006, March). Developmental Outcomes Associated with the After-School Contexts of Low-Income Children and Youth: A study of Promising after-school programs and organized activities.

Vandell, D. L., Pierce, K. M., O'Cadiz, P., Hall, V., Karsh, A., Westover, T. (2010). Youth Outcome Measures Online Toolbox. Unpublished manuscript, Department of Education, University of California – Irvine.

Vandell, D. L., Reisner, E. R., Pierce, K. M., Brown, B. B., Lee, D., Bolt, D., & Pechman, E. M. (2006). *The Study of Promising After-School Programs: Examination of Longer Term Outcomes After Two Years of Program Experiences*. Retrieved from University of Wisconsin, Wisconsin Center for Education Research website: [http://childcare.wceruw.org/pdf/pp/year\\_3\\_report\\_final.pdf](http://childcare.wceruw.org/pdf/pp/year_3_report_final.pdf).

Vandell, D. L., Reisner, E. R., Brown, B. B., Dadisman, K., Pierce, K. M., Lee, D., & Pechman, E. M. (2005). *The Study of Promising After-School Programs: Examination of Intermediate Outcomes in Year 2*. Report to the Charles Stewart Mott Foundation. Retrieved from <http://www.education.uci.edu/childcare/pdf/afterschool/PP%20Examination%20in%20Year%202.pdf>.

