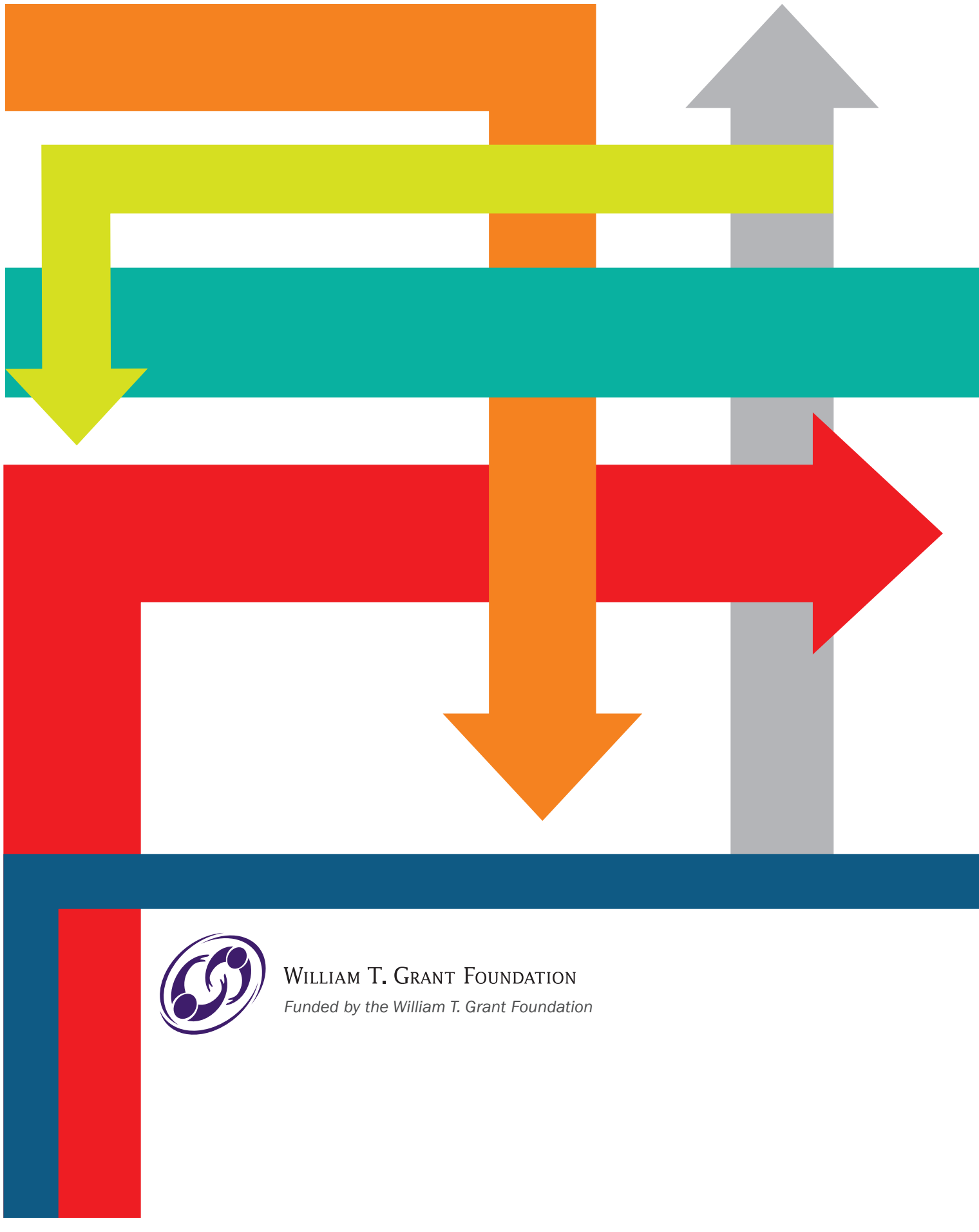
the forum
FOR YOUTH INVESTMENT
Published by The Forum for Youth Investment
September 2011

From Soft Skills to Hard Data:

MEASURING YOUTH PROGRAM OUTCOMES

Alicia Wilson-Ahlstrom & Nicole Yohalem, The Forum for Youth Investment;
David DuBois, University of Illinois at Chicago &
Peter Ji, Adler School of Professional Psychology



William T. Grant Foundation

Funded by the William T. Grant Foundation

Table of Contents

Overview and Purpose

Youth programs operating during the non-school hours are important partners that work alongside families and schools to support learning and development. Some programs prioritize academics; others prioritize enrichment, recreation or leadership development; others weave together a combination of these. Whether focused on sports, art or community service, most of these programs aim to develop cross-cutting skills that will help young people be successful now and help ensure they are ready for college, work and life.

Helping to build what are often referred to as "social-emotional" or "21st century skills" is an important contribution that many youth programs make and more could be making. Yet these efforts remain underrepresented in the program evaluation literature, in part because they cannot be measured using administrative records or other databases to which schools and programs might have easy access.

Practitioners and funders regularly ask us for advice about how to measure these skills. In response we developed this guide, which summarizes information about tools that programs can use to measure youth progress in these areas. The guide builds on and complements several related resources available in the field (for a listing, see *Other Collections of Youth Outcome Measures*, page 5).

Our goal is to help practitioners choose conceptually grounded and psychometrically strong measures of important skills and dispositions that cut across academic achievement and other distal youth outcomes like risk behavior, mental health and employment. We also hope to encourage the development of additional measures in areas where our review reveals gaps. In a time of increasing pressure on programs to improve policy-relevant outcomes, we want to facilitate access to good measurement tools. This can help advance the out-of-school time (OST) field and facilitate collaboration among practitioners working toward common goals, both in school and out.

Why these Outcome Areas?

Although consensus has yet to emerge about what to call these skills, there is growing recognition that they are critically important. *Preparing Students for College and Careers,* one of the most recent among many policy research efforts on this subject, notes that "according to teachers, parents, students and Fortune 1000 executives, the critical components of being college- and career-ready focus more on higher-order thinking and performance skills than knowledge of challenging content."[i] Over 400 employers surveyed in 2006 identified collaboration, work ethic and communication as among the most important skills necessary to succeed in the workplace. Yet only 24 percent of employers believe that new employees with four-year college degrees have "excellent" applied skills in these areas.[ii]

The policy momentum building in this area is notable, but we decided to review measures of these skills for several additional reasons. First, research suggests these are important to school and workplace success as well as to risk behavior reduction.[iii] Also, the literature suggests that when programs achieve impacts in these areas, they also make progress on more traditional academic measures like grades and test scores.[iv] And despite growing interest, efforts to measure these areas effectively are still evolving.[v]

We also believe these outcome areas represent a strategic niche or, in economic terms, a "comparative advantage" for many youth programs. OST programs operate with limited resources yet have significant flexibility compared with schools. They can play a powerful role in building skills that matter for learning and development. But to live up to this potential, activities need to align with outcomes, and programs need tools that are accessible and that adequately measure the skills and dispositions that they expect young people to develop. Not surprisingly, experts from the OST field encouraged us to focus on these skills during the planning stages of this project.

We arrived at four specific skill areas to focus on – communication, relationships and collaboration, critical thinking and decision making, and initiative and self-direction – by reviewing commonly cited frameworks developed by the Collaborative for Academic, Social and Emotional Learning (CASEL), the Partnership for 21st Century Skills and the U.S. Department of Labor.[vi] In addition to identifying common constructs across these frameworks, we decided to focus on specific, skill- and ability-oriented outcomes and to prioritize skill areas that are amenable to intervention by OST programs. We also focused on skills that are cross-cutting, which means we left out some skills that relate to specific content knowledge (e.g., technology and global awareness).

Other Collections of Youth Outcome Measures

ToolFind, United Way of Mass Bay with NIOST
www.toolfind.org

Compendium of Assessment and Research Tools (CART), RMC Research Corporation
http://cart.rmcdenver.com

Measurement Tools for Evaluating Out-of-School Time Programs, Harvard Family Research Project
www.hfrp.org/out-of-school-time/publications-resources

Tools for Research & Evaluation of Intervention Programs, Outdoor Education R&D Center
http://wilderom.com/tools.html

Assessment Tools in Informal Science, PAER at Harvard University, in collaboration with 4-H
www.pearweb.org/atis

Supporting Evaluation and Research Capacity Hub website, CYFAR/USDA
https://cyfernetsearch.org/

Compendium of Measures Used in P-12 Evaluations of Educational Interventions, IES and Mathematica
http://ies.ed.gov/ncee/pubs/20104012/pdf/20104013.pdf

Online Evaluation Resource Library (OERL), SRI International
http://oerl.sri.com

Youth Outcomes Compendium, Child Trends
www.childtrends.org/what_works/clarkwww/compendium_intro.asp

Compendium of Preschool - Elementary School SEL and Associated Assessment Measures, CASEL
http://casel.org/wp-content/uploads/2011/04/Compendium_SELTools.pdf

Afterschool Youth Outcomes Inventory, PASE
www.pasesetter.com/documents/pdf/Outcomes/OutcomesInventory_8Nov10%20FINAL.pdf

SEL Measures for Middle School Youth, UW Social Development Research Group for Raikes Foundation
http://raikesfoundation.org/Documents/SELTools.pdf

Measuring Student Engagement in Upper Elementary Through High School, REL Southeast
http://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_2011098_sum.pdf

By no means do we suggest that this is a comprehensive list of important skills and dispositions, or that these are the only skills that OST programs should focus on or measure. For example, many programs track academic outcomes like school attendance, homework completion, grades or standardized test scores. However, they typically track these outcomes using data obtained from school records, which means program leaders rarely face decisions about what instrument to use.

Finally, our decision to focus on these four areas was also a practical one. Limiting the number of tools allowed us to conduct detailed reviews and helped ensure that this resource would build on rather than be redundant with other resources in the field.

Skill Areas Featured in this Report

Communication: Self-expression, listening, public speaking and recognizing non-verbal cues.

Relationships & Collaboration: Interpersonal skills, team work, flexibility and cultural competence.

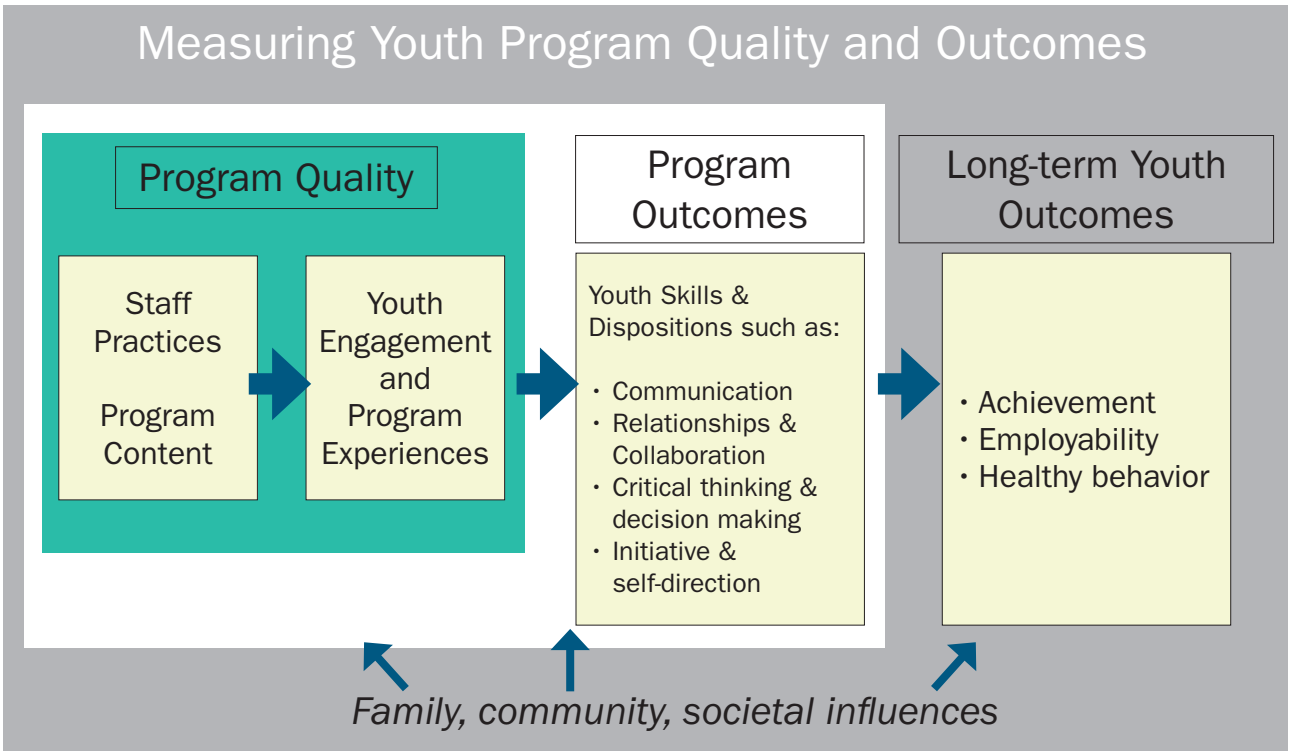
Critical Thinking & Decision-making: Reasoning, making judgments and decisions, responsible problem-solving, creativity and accessing, evaluating, and using information.

Initiative & Self-direction: Self-awareness, setting and working toward goals, self-management, working independently, and guiding and leading others.

Why these Instruments?

In determining what instruments to include (see Table 1 for a list) we considered several factors. Before describing those factors, we should explain why we focused on measures of youth outcomes as opposed to program process or quality.

Figure 1: Adapted from the David P. Weikart Center for Youth Program Quality



In 2007 we published *Measuring Youth Program Quality*[vii], which reviewed observational measures of youth program practices. Although we remain strongly committed to assessing the quality of program practices – especially interactions among youth and adults at the "point-of-service" – it is critical that improvements in program practices lead to good outcomes for participants. Because many programs are trying to measure outcomes, we developed this guide as a companion document to our 2007 work on practices. Here we looked for ways for programs to assess whether particular skills or dispositions transfer outside of the program

setting (although some instruments include items or scales focused on the extent to which youth use specific skills in the program itself). Figure 1 (on the prior page) shows how the outcome measures reviewed here fit into a broad theory of change about youth program impact.

In selecting outcome measures to review, we first identified measures where a majority of the content (more than half of the items in a given scale) mapped directly onto one of our four areas of interest: communication, relationships and collaboration, critical thinking and decision making, and initiative and self-direction.

We looked for measures that were appropriate for use in a range of settings, including OST programs, schools, youth development organizations and camps. We included some measures that have not been used extensively in OST settings but could be. Our focus was on programs serving upper elementary- through high school-age youth, a decision driven in part by the significant work already done to review measures appropriate for use with younger children.[viii] We also prioritized measures that are accessible and relatively low-burden for practitioners to implement.

On the technical side, we looked for instruments that had been investigated for scale reliability, factor structure and sensitivity to OST program impact. That decision led to the exclusion of some promising tools that are early in their development, but reflects our commitment to ensuring that practitioners have access to instruments that yield valid and reliable information. We did include some measures that did not meet all of our technical criteria in cases where a measure is already used extensively in OST programs and validation efforts are ongoing. We hope the criteria that guided our technical review *(see Framework and Criteria for Ratings of Reliability and Validity Evidence, p. 61)* provide a useful roadmap for further testing and development of instruments that are not included here.

Table 1: Instruments, Developers and Availability

Instrument	Developer	Website
California Healthy Kids Survey Resilience & Youth Development Module (RYDM)	Greg Austin and Mark Duerr, WestEd	http://chks.wested.org/
Developmental Assets Profile (DAP)	Search Institute	www.search-institute.org/survey-services/surveys/developmental-assets-profile
Devereaux Student Strengths Assessment (DESSA)	Devereux Center for Resilient Children	www.k5kaplan.com
San Francisco Beacons Survey	Public/Private Ventures (P/PV)	http://www.ppv.org/ppv/publication.asp?search_id=5&publication_id=168&section_id=0
Social Skills Improvement System (SSIS)	Frank Gresham and Stephen Elliott, Pearson	www.pearsonassessments.com/HAIWEB/Cultures/enus/Productdetail.htm?Pid=PAa3400&Mode=summary
Survey of Afterschool Youth Outcomes (SAYO)	Wendy Surr and Allison Tracy, National Institute on Out-of-School Time (NIOST)	www.niost.org/content/view/1653/282/
Youth Outcomes Battery	Jim Sibthorp and Dr. Gary Ellis, American Camp Association (ACA)	www.acacamps.org/research/enhance/youth-outcomes-resources
Youth Outcome Measures Online Toolbox	Deborah Lowe Vandell, Kim Pierce, Pilar O'Cadiz, Valerie Hall, Andrea Karsh, and Teresa Westover	http://childcare.wceruw.org/form3.html

Using the Guide

While programs collect outcome data for a variety of reasons – including the desire to better fit program activities to the needs of young people, the desire to assess how much a program is improving outcomes and the dictates of funders – several considerations are critical to selecting a measurement tool.

First and foremost, outcome measures should reflect the goals and activities of the program. Programs should measure outcomes that they value and that they are intentionally trying to influence. Second, programs should use measures that will yield valid and reliable information. Finally, programs should also consider a host of important practical issues such as the cost, ease of administration and accessibility of the tools. This guide includes information on all of these considerations.

For each instrument, we summarize the origins and focus of the tool, include sample items and discuss user and technical considerations. Where possible, information is provided about length, cost, format (e.g., Web vs. paper; translations), supplemental measures and tools, and training (whether it is available or required). Our technical reviews focus on the degree to which reliability and validity have been established. Reliability speaks to whether an instrument yields consistent information, while validity speaks to whether a particular instrument in fact measures what it intends to measure.

We summarize the technical properties of each instrument as a whole and provide more detailed reviews of the scales within each instrument that map most directly onto the four skill areas that are discussed above. For each relevant scale we rate the strength of evidence for reliability and validity — the former derived from consideration of internal consistency, inter-rater and test-retest reliability; the latter from consideration of convergent, discriminant, criterion and construct validity. For a discussion of the importance of psychometrics and definitions of all of these terms, (see *Psychometrics: What are they and why are they useful?, p.51).* For those readers who are interested in detailed analyses of reliability and validity evidence for each scale and want to understand the process used to arrive at technical ratings, please see the Technical Appendix.

The technical ratings should by no means be considered final. In most cases, the instrument developers are continually gathering evidence of reliability and validity. Readers are encouraged to ask developers for updated information and watch for forthcoming updates to this report.

Finally, a word of caution: We have tried to identify useful measures that are psychometrically sound so that if change is detected, users can be confident that change is in fact occurring. But attribution – or determining whether that change is a function of a specific program – requires specific approaches to study design that are beyond the scope of this report.

Looking across the Instruments

This section includes some observations about this set of eight instruments as a whole, and several summary charts. The section that follows provides detailed information about each instrument.

What skills do these instruments measure?

All eight of the instruments include at least one scale that addresses collaboration and relationships and initiative and self-direction. Despite the fact that many youth programs focus on building critical thinking and decision-making skills, fewer than half of the instruments reviewed measure these outcomes, and only two have scales that measure communication skills. It is important to note that all of the instruments also measure constructs that fall outside of the four areas we focused on. See Table 2 for a full listing of skills assessed by each instrument and Table 3 for a listing of scales by skill area.

How accessible and user-friendly are these instruments?

Only three of the eight measures are currently available free of charge; others have associated costs ranging from nominal one-time fees to more substantial per-survey costs. While user manuals and related resources are available in most cases, specific user training is available (for a fee) for four of the eight instruments.

Tables with normative data designed to facilitate comparison of youth in a given program to a larger population are available in four cases, although several developers are working to make such data available. See Tables 4 and 5 for a summary of these and other user considerations.

To what extent have reliability and validity been established?

There is evidence that the scales on each of the eight instruments generate consistent responses, or are reliable. However the strength of reliability evidence varies across the eight instruments and typically across scales within each individual instrument (see Table 6), as does the extent to which reliability has been established for different groups (e.g. age, gender and ethnicity). For all eight of the instruments included in the guide, there is some evidence that the scales measure what they intend to measure, or are valid. However, the strength of validity evidence varies across the eight instruments and typically across the scales within each individual instrument (see Table 6).

From a technical standpoint, what additional information would be useful?

As the developers and other scholars continue to work with these instruments, there are several areas where additional information would be useful, particularly in terms of advancing validation efforts. For example, additional work on convergent and discriminant validity, or the extent to which scales in fact measure their specific intended constructs, would be useful for all eight instruments. Additional efforts to assess the degree to which scores on scales relate in expected ways to relevant criterion or outcome measures, obtained either at the same time (concurrent validity) or at some point in the future (predictive validity), would also be helpful in all cases. Finally, for most instruments, efforts to assess how useful scales are in detecting effects of OST participation would help advance the field.

Table 2: Skill Areas Assessed

Instrument	Communication	Relationships & Collaboration	Critical Thinking & Decision-making	Initiative & Self-Direction	What Else Does it Measure?
California Healthy Kids Survey Resilience & Youth Development Module (RYDM)		X	X	X	Caring Relationships[1]; High Expectations; Meaningful Participation; Goals and Aspirations; School Connectedness
Developmental Assets Profile (DAP)		X		X	Support; Empowerment; Boundaries and Expectations; Constructive Use of Time; Positive Values
Devereaux Student Strengths Assessment (DESSA)		X	X	X	Optimistic Thinking
San Francisco Beacons Survey		X		X	Passive Reaction to Social Challenge; Non-Familial Support; Peer Support; Adult Support at the Beacons; Variety of Interesting Activities offered at the Beacons
Social Skills Improvement System (SSIS)	X	X		X	Cooperation; Responsibility; Competing Problem Behaviors; Academic Competence
Survey of Afterschool Youth Outcomes (SAYO)	X	X	X	X	Engagement in Learning; Homework; Academic Performance; Program Experiences; Environment; Sense of Competence as a Learner[2]; Future Planning and Expectations
Youth Outcomes Battery	X	X	X	X	Family Citizenship; Perceived Competence; Affinity for Nature; Spiritual Well-being; Camp Connectedness
Youth Outcome Measures Online Toolbox		X	X	X	Aggressive Behavior with Peers; Academic Performance; Misconduct; Reading/English Efficacy; Math Efficacy

Note: An X in a box means the instrument includes a scale where more than half of the scale's items map directly onto the construct in question.

[1] Caring Relationships, High Expectations, and Meaningful Participation each contain items that measure these in school, community, home and/or peer support contexts.
[2] Measure includes a sense of competence in reading, writing, math and science.

Table 3: Scales Organized by Skill Areas

Instrument	Communication	Relationships & Collaboration	Critical Thinking & Decision-making	Initiative & Self-Direction
California Healthy Kids Survey Resilience & Youth Development Module (RYDM)		Empathy; Cooperation & Communication	Problem Solving	Self-Awareness; Self-Efficacy
Developmental Assets Profile (DAP)		Social Competencies		Commitment to Learning; Positive Identity
Devereaux Student Strengths Assessment (DESSA)		Social Awareness; Relationship Skills; Self-Management	Decision Making	Personal Responsibility; Goal-Directed Behavior; Self-Awareness
San Francisco Beacons Survey		Positive Reaction to Social Challenge		School Effort; Self-Efficacy; Leadership; Time Spent in Challenging Learning Activities
Social Skills Improvement System (SSIS)	Communication	Assertion; Empathy; Engagement; Self-Control		
Survey of Afterschool Youth Outcomes (SAYO)	Communication Skills	Sense of Competence Socially; Relations with Adults; Relations with Peers	Problem-Solving Skills	Behavior in the Classroom; Initiative; Future Planning – My Actions
Youth Outcomes Battery		Friendship Skills; Teamwork	Problem Solving Confidence	Independence; Interest in Exploration; Responsibility
Youth Outcome Measures Online Toolbox		Prosocial Behavior; Social Skills; Social Competencies		Work Habits; Task Persistence

Note: This does not include all of the scales from each instrument, only those that map onto the skill areas that are the focus of this guide.

Table 4: User Considerations in Selecting Measures – Populations and Settings

Measures	Target Age/ Grades	Settings Tool has Been Tested In	Availability of Normative Data
California Healthy Kids Survey Resilience & Youth Development Module (RYDM)	Middle & High School	Primarily Schools	Data collected and analyzed on large numbers of California youth who have taken the *Resiliency & Youth Development Module*. Reports summarizing these data and descriptive information about the state-level sample are available.
Developmental Assets Profile (DAP)	Middle & High School	OST programs; Schools; therapeutic settings	Normative data designed to facilitate comparison of youth in a given program to a larger population are not available at this time.
Devereaux Student Strengths Assessment (DESSA)	K – 8	Schools; Residential programs; Clinical settings	Normative data are available for each scale of the DESSA; based on a standardization sample consisting of nearly 2,500 children that sample is reported to closely approximate the K-8 population of the U.S. with respect to age, gender, geographic region of residence, race/ethnicity, and socioeconomic status based on data published in 2008 by the U.S. Census Bureau. Norm reference cards are available for purchase and are included in the DESSA kit.
San Francisco Beacons Survey	Middle School	Beacons afterschool programs	Normative data designed to facilitate comparison of youth in a given program to a larger population are not available at this time.
Social Skills Improvement System (SSIS)	Elementary – High School	Primarily schools; Clinical settings	Tested on a normative sample of 4,700 youth ages 3-18. In addition, 385 teachers and 2,800 parents provided ratings. Sampling was conducted on a national standardization sample aligned with the demographic results of the 2006 U.S. Census. Sampling was conducted on a national standardization sample aligned with the demographic data published by the 2006 U.S. Census Bureau. Information about using norms is included in kits.
Survey of Afterschool Youth Outcomes (SAYO)	4th – 8th; 9th – 12th	OST programs/Afterschool programs	Normative data designed to facilitate comparison of youth in a given program to a larger population are not available at this time.
Youth Outcomes Battery	Middle & High School	Primarily camps (both day and residential)	ACA recently began collecting normative data on the Basic version of the *Youth Outcomes Battery*. These data are intended to allow individual camps to compare their scores with representative scores from typical ACA camps. (Data offer limited comparison value for non-residential camp programs because 75% were collected on residential camps.) Details related to gender, age, race/ethnicity and day/resident programming are forthcoming. Guidance on how to use norms for comparison purposes is available at www.acacamps.org/research/enhance/youth-outcomes-resources/norms.
Youth Outcome Measures Online Toolbox	Middle School	Middle school OST programs	Normative data designed to facilitate comparison of youth in a given program to a larger population are not available at this time.

Table 5: User Considerations in Selecting Measures – Accessibility and Supports

Instrument	Approx. Time to Complete	Cost	Training Available	Companion/Related Tools	Additional Information & Supports
California Healthy Kids Survey Resilience & Youth Development Module (RYDM)	~40 minutes	Free	Upon request	Part of the California School Climate, Health and Learning survey tools. Includes a School Climate survey and Parent survey	- Interested programs should contact the California DOE for permission to use - Guidebook available online - Modifications needed to use for individual program evaluation purposes - Survey can be customized; a database of sample questions used is available
Developmental Assets Profile (DAP)	~20 minutes[3]	$195 for 50 surveys/ scoring sheets	No	Developmental Assets Community Mobilization ("40 Assets") survey	- Survey available online or paper copy - User's guide included
Devereaux Student Strengths Assessment (DESSA)	N/A	$115.95 for standard kit, including user manual and forms. $39.95 for 25 additional forms	Yes	*DESSA-Mini*	- Programs seeking more information prior to purchase may read an introduction to the tool - Fee-based in-service training available but not required - Free video and audio presentations also available
San Francisco Beacons Survey	~35 minutes	Free	No	*Youth Feedback Form* (on program experiences)	- Interested programs should contact the developer for access to and guidance on the survey
Social Skills Improvement System (SSIS)	~25 minutes	$248.45 for starter kit, including rating scales and manual ($517.35 for computer-scored kit). $43.05 for 25 hand-scored surveys; $53.60 for 25 computer-entry surveys	No	Part of the *Social Skills Improvement System* which includes guides for Performance Screening and Improvement Planning	- ASSIST software provides computer scoring and reporting, including individual, progress and multi-rater reports - Online direct links to suggested interventions with the *SSIS* Intervention Guide - Available in Spanish
Survey of Afterschool Youth Outcomes (SAYO)	~20 minutes	$250 for unlimited one year site license	Yes	Part of the *APAS* assessment system which includes an observational tool for assessing quality	- Youth surveys available online only - Training available in-person or online - Survey may be customized
Youth Outcomes Battery	N/A	$5 (members) or $15 (non-members) per scale	No	Can be used in tandem with an 8-step program evaluation process	- Designed with camps in mind, though "camp" language can be replaced with "program" - Guidelines available online
Youth Outcome Measures Online Toolbox	~25 minutes	Varies based on number of sites, number of students per site, and level of analyses	Upon request	Teacher Student Report, Program Staff Student Report, Program Observation tool and elementary level survey	- Interested programs should contact the developer for access and guidance on the survey

[3] Time based on recommended survey length of no more than 50 questions selected from a menu of scales.

Table 6: Technical Properties Summary

	Reliability			Validity	
	Is there evidence that the scales on the instrument generate consistent responses?	How strong is available reliability evidence?	Reliable for what groups?	Is there evidence that the scales on the instrument are good measures of what they intend to measure?	How strong is available validity evidence?
California Healthy Kids Survey Resilience & Youth Development Module (RYDM)	Yes	Moderate-to-Substantial	Students in grades 7, 9 and 11; male and female youth; youth belonging to different racial/ethnic groups	Yes	Moderate
Developmental Assets Profile (DAP)	Yes	Substantial	Middle and high school students; male and female youth; youth from different racial/ethnic groups	Yes	Moderate
Devereaux Student Strengths Assessment (DESSA)	Yes	Moderate	Elementary school students	Yes	Limited-to-Moderate
San Francisco Beacons Survey	Yes	Limited-to-Moderate	Primarily for middle school aged youth	Yes	Moderate
Social Skills Improvement System (SSIS)	Yes	Moderate-to-Substantial	Male and female youth ages 12 and under and ages 13-18	Yes	Moderate
Survey of Afterschool Youth Outcomes (SAYO)	Yes	Substantial	Elementary/middle and high school students; male and female youth; youth from different racial/ethnic groups	Yes	Moderate-to-Substantial
Youth Outcomes Battery	Yes	Limited	Reliability findings have not been reported for specific groups of youth	Yes	Limited
Youth Outcome Measures Online Toolbox	Yes	Substantial	Elementary and middle school students; male and female youth; English Language Learner youth; youth from different racial/ethnic groups	Yes	Moderate

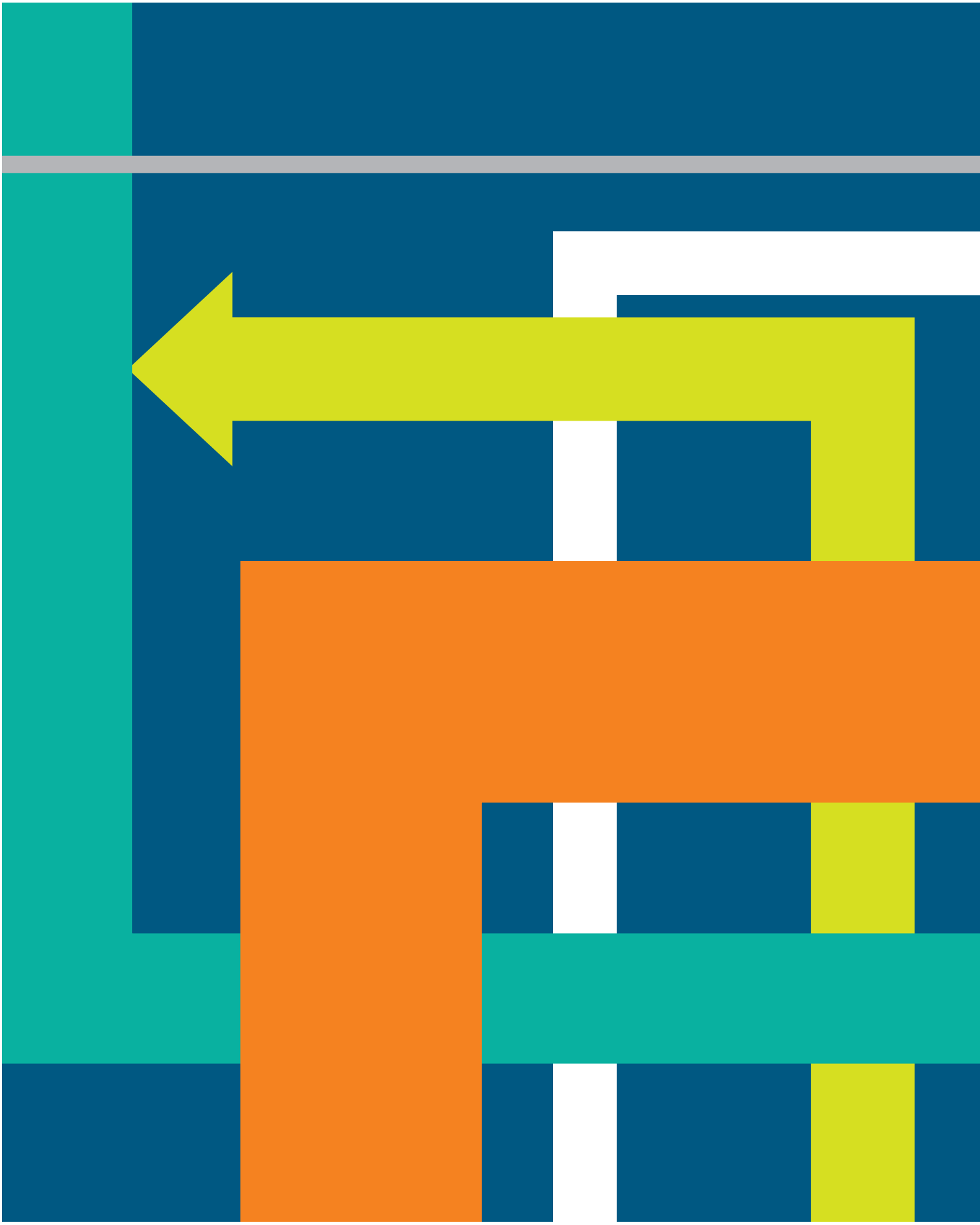
Note: For detailed explanation of our rating scale for reliability and validity evidence and how we arrived at ratings for Tables 6 - 10, see *Framework and Criteria for Ratings of Reliability and Validity Evidence* on p. 62. The range of rating levels include *None, Limited, Moderate, Substantial, and Extensive.*

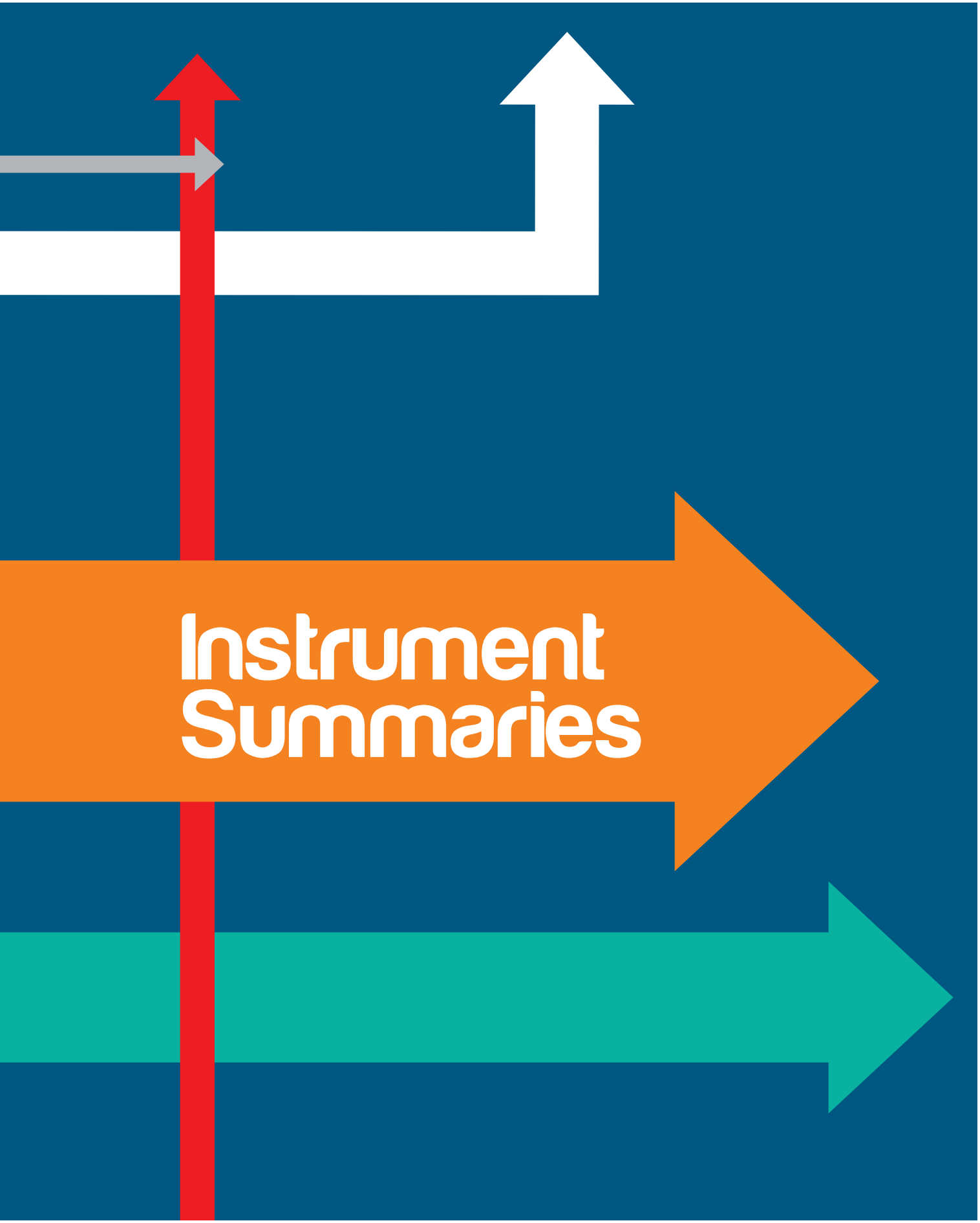
Table 7: Relationships & Collaboration Scales: Technical Properties Summary		
	Overall Reliability Rating	**Overall Validity Rating**
Empathy *(RYDM)*	Moderate-to-Substantial	Moderate
Cooperation and Communication *(RYDM)*	Moderate	None-to-Limited
Social Competencies *(DAP)*	Substantial	Limited-to-Moderate
Social Awareness *(DESSA)*	Moderate	Limited-to-Moderate
Relationship Skills *(DESSA)*	Moderate	Moderate
Self-Management *(DESSA)*	Moderate	Moderate
Positive Reaction to Social Challenge *(Beacons)*	Limited	Moderate
Assertion - Teacher *(SSIS)*	Moderate	Moderate
Assertion - Student *(SSIS)*	Moderate-to-Substantial	Moderate
Empathy - Teacher *(SSIS)*	Moderate	Moderate
Empathy - Student *(SSIS)*	Moderate-to-Substantial	Moderate
Engagement - Teacher *(SSIS)*	Moderate-to-Substantial	Moderate
Engagement - Student *(SSIS)*	Moderate-to-Substantial	Moderate-to-Substantial
Self-Control - Teacher *(SSIS)*	Moderate-to-Substantial	Moderate
Self-Control - Student *(SSIS)*	Moderate-to-Substantial	Moderate
Sense of Competence Socially *(SAYO)*	Substantial	Moderate-to-Substantial
Relations with Adults *(SAYO)*	Substantial	Moderate-to-Substantial
Relations with Peers *(SAYO)*	Substantial	Moderate-to-Substantial
Friendship Skills *(Youth Outcomes Battery)*	Limited	Limited
Teamwork *(Youth Outcomes Battery)*	Limited	Limited
Prosocial Behavior *(Online Toolbox)*	Substantial	Moderate-to-Substantial
Social Skills *(Online Toolbox)*	Substantial	Moderate
Social Competencies *(Online Toolbox)*	Moderate-to-Substantial	Moderate

Table 8: Initiative & Self-Direction Scales – Technical Properties Summary		
	Overall Reliability Rating	**Overall Validity Rating**
Self-Awareness *(RYDM)*	Substantial	Moderate
Self-Efficacy *(RYDM)*	Moderate	Limited-to-Moderate
Commitment to Learning *(DAP)*	Substantial	Limited-to-Moderate
Positive Identity *(DAP)*	Substantial	Moderate
Personal Responsibility *(DESSA)*	Moderate	Limited-to-Moderate
Goal-Directed Behavior *(DESSA)*	Moderate	Limited-to-Moderate
Self-Awareness *(DESSA)*	Moderate	Limited-to-Moderate
School Effort *(Beacons)*	Limited	Moderate
Self-Efficacy *(Beacons)*	None	Limited-to-Moderate
Leadership *(Beacons)*	None	None-to-Limited
Time Spent in Challenging Learning Activities *(Beacons)*	None	Limited
Behavior in the Classroom *(SAYO)*	Moderate-to-Substantial	Moderate-to-Substantial
Initiative *(SAYO)*	Substantial	Moderate-to-Substantial
Future Planning – My Actions *(SAYO)*	Substantial	Moderate-to-Substantial
Independence *(Youth Outcomes Battery)*	Limited	Limited
Interest in Exploration *(Youth Outcomes Battery)*	Limited	Limited
Responsibility *(Youth Outcomes Battery)*	Limited	None-to-Limited
Work Habits *(YO Toolbox)*	Moderate-to-Substantial	Limited-to-Moderate
Task Persistence *(YO Toolbox)*	Substantial	Limited-to-Moderate
Social Competencies *(YO Toolbox)*	Moderate-to-Substantial	Moderate

Table 9: Communication Scales – Technical Properties Summary		
	Overall Reliability Rating	**Overall Validity Rating**
Communication - Teacher *(SSIS)*	Moderate-to-Substantial	Moderate
Communication - Student *(SSIS)*	Moderate-to-Substantial	Moderate
Communication Skills *(SAYO)*	Substantial	Moderate-to-Substantial

Table 10: Critical Thinking & Decision-Making Scales – Technical Properties Summary		
	Overall Reliability Rating	**Overall Validity Rating**
Problem Solving *(RYDM)*	Moderate	Limited-to-Moderate
Decision Making *(DESSA)*	Moderate	Limited-to-Moderate
Problem-Solving Skills *(SAYO)*	Substantial	Moderate-to-Substantial
Problem-Solving Confidence *(Youth Outcomes Battery)*	Limited	Limited





California Healthy Kids Survey Resilience & Youth Development Module

Overview and Purpose

The *California Healthy Kids Survey (CHKS)* is a statewide survey administered to students in grades 5-8 enrolled in California. The purposes of *CHKS* include helping schools monitor and address mental and physical health needs (especially as they affect academic performance), improving school climate and learning supports, and increasing the quality of health, prevention and youth development programs. *CHKS* was developed by WestEd for the California Department of Education. Until the 2010-2011 school year, California school districts that accepted funds under Title IV were required to administer the *CHKS*. In recent years, schools and youth programs from other parts of the country have used and adapted the survey. In addition to the core survey, 11 supplemental modules can be used to customize the survey to meet local needs. The *Resilience & Youth Development Module (RYDM)* aligns most closely with our focus and therefore is the subject of this review. The *RYDM* is based on a conceptual framework that links environmental and internal resilience assets to improved health, social and academic outcomes.

Content

There are middle school and high school versions of the *RYDM*; each includes a shorter and longer form, with 33 and 56 questions respectively. The full version includes scales that assess home and peer environments that are not included in the shorter version.

Each question (see sample items) follows a four-point response scale: not at all true, a little true, pretty much true, very much true. To assist with interpretation of a youth's scores on each scale, guidelines are available for categorizing scores as high, moderate or low. Scale scores (average item response) over 3 are categorized as "high", those between 2 and 3 are categorized as "moderate", and those less than 2 are categorized as "low." Programs may find it useful to report percentages of students whose scores fall in the high, moderate or low categories for each scale.

The *RYDM* includes the following scales:

- Caring Relationships (Includes four scales: Community Caring Relationships, School Caring Relationships, Home Caring Relationships, Peer Caring Relationships)
- High Expectations (Includes four scales: Community High Expectations, School High Expectations, Home High Expectations, Peer High Expectations)
- Meaningful Participation (Includes four scales: Community Meaningful Participation, School Meaningful Participation, Home Meaningful Participation, Peer Meaningful Participation)
- Cooperation and Communication*
- Empathy*
- Problem Solving*
- Goals and Aspirations
- Self-Awareness*
- School Connectedness
- Self-Efficacy*

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

Sample Items from *CHKS Resilience & Youth Development Module* Scales Reviewed in this Guide

I can work with someone who has different opinions than mine.
(Cooperation and Communication)

I feel bad when someone gets their feelings hurt.
(Empathy)

I try to work out problems by talking or writing about them.
(Problem Solving)

I understand my moods and feelings.
(Self-Awareness)

I can do most things if I try.
(Self-Efficacy)

User Considerations

In this section we discuss several considerations related to the *RYDM*, including availability of normative data, accessibility, ease of use and available supports.

Accessibility

The *RYDM* and related *CHKS* instruments are available for free on the *California Healthy Kids Survey* website and can be used with permission from the California Department of Education.

Ease of Use

The *RYDM* uses a paper/pencil format. A typical youth will finish the survey in under 20 minutes. The website provides instructions for administering the survey.

Availability of Norms

Normative data that characterize what is usual within a defined population can help programs better understand the populations they serve and the effects of their programs. The administrators of the *California Healthy Kids Survey* have collected and analyzed data on large numbers of California youth who have taken the *RYDM*. Reports summarizing these data are available on http://chks.wested.org/reports and descriptive information about the state-level sample is provided in this report: http://chks.wested.org/resources/Secondary_State_0709_Main.pdf.

Available Supports

WestEd provides training and analysis support to programs outside of California on a cost recovery basis. They also have a range of resources on their website, including background information on the framework on which the instruments are based, guidelines for customizing and administering the survey, and information on interpreting and reporting scores.

Technical Properties

This section provides information about the overall technical properties of the *RYDM* and of specific scales that map onto the skill areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for those five scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the CHKS Resilience & Youth Development Module

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes

2. For what groups?

- Students in grades 7, 9 and 11
- Male and female youth
- From different racial/ethnic groups (White, African-American, Mexican-American, Chinese-American)

3. How strong is available reliability evidence?

- Moderate-to-Substantial

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Moderate

6. What is the nature of that evidence?

- Exploratory and confirmatory factor analysis support for viewing scales on the RYDM as measures of distinct environmental and personal resilience assets.
- Significant associations of RYDM scale scores in expected directions with youth self-reports of substance use, violence, psychological well-being and school adjustment (grades, truancy).

7. What are some of the questions that would be useful for scholars to address as they continue to work with this instrument?

- To what extent do RYDM scales measure their specific intended constructs – e.g., does the Problem-Solving scale correlate with other established measures of skills in this area and less so with measures of other types of skills?
- What are the cumulative and unique contributions of RYDM scales, when considered collectively, to the prediction of different types of youth outcomes?
- To what extent do RYDM scales predict outcomes at later points in a youth's schooling or development?
- What is the sensitivity of RYDM scales for detecting expected effects of OST program participation?

Reliability and Validity of *CHKS Resilience & Youth Development Module* Scales Reviewed in this Guide

Scale	Number of Items	Evidence of Reliability	Evidence of Validity	Corresponding Skill Area in this Guide
Cooperation and Communication	3	Moderate	None-to-Limited	Relationships & Collaboration
Empathy	3	Moderate-to-Substantial	Moderate	Relationships & Collaboration
Problem Solving	3	Moderate	Limited-to-Moderate	Critical Thinking & Decision Making
Self-Awareness	3	Substantial	Moderate	Initiative & Self-Direction
Self-Efficacy	3	Moderate	Limited-to-Moderate	Initiative & Self-Direction

For More Information

T. Kiku Annon
Research Associate, WestEd
Regional Coordinator, CA School Climate, Health & Learning Survey
4665 Lampson Ave.
Los Alamitos, CA 90720
kannon@wested.org
(562) 799-5127 (Phone)
(562) 799-5151 (Fax)

Developmental Assets Profile

Overview and Purpose

The *Developmental Assets Profile (DAP)* was developed by Search Institute in 2004. Based on the Institute's developmental assets framework, the *DAP* measures the external assets (relationships and opportunities provided by others) and internal assets (values, skills and self-perceptions) of youth in grades 6-12. Search Institute developed the *DAP* in response to numerous requests for a measure of developmental assets appropriate for program evaluation and clinical purposes. It can be used to assess individual youth or as a group assessment for all participants in a program.

Content

The *DAP* is a 58-item self-report questionnaire. Youth are asked how true each statement is for them in the context of a three-month time frame and respond using a four-point scale: not at all, rarely/somewhat, sometimes/very often, extremely/almost always.

The *DAP* can be scored to reflect the types and degree of developmental assets that each youth reports in each of the following categories:

- Support
- Empowerment
- Boundaries and Expectations
- Constructive Use of Time
- Commitment to learning*
- Positive Values
- Social Competencies*
- Positive Identity*

Alternatively, items can be re-grouped to yield scores reflecting assets associated with each of the following developmental contexts: personal, social, family, school and community.

Sample Items from *DAP* Scales Reviewed in this Guide

I am actively engaged in learning new things. (Commitment to Learning)

I build friendships with other people. (Social Competencies)

I am developing a sense of purpose in my life. (Positive Identity)

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

User Considerations

This section discusses the *DAP* in terms of several important user considerations, including accessibility, ease of use, availability of normative data and other supports available to users.

Accessibility

The *DAP* may be administered online or in a paper/pencil format. A basic package of 50 surveys (online or paper/pencil), 50 self-scoring profile forms and the user manual costs $195. Additional surveys and forms may be purchased.

Ease of Use

Search Institute suggests it takes a typical youth 10-15 minutes to complete the *DAP*. The survey is self-explanatory and requires no special training to administer. A Web-based scoring platform (included in the survey package) allows users to administer, score, view, print and export *DAP* results. Materials and procedures for hand-scoring are also available.

Availability of Norms

Normative data that characterize what is usual within a defined population can help programs better understand the populations they serve and the effects of their programs. Although norms based on a representative national sample of youth are not yet available for the *DAP*, Search Institute is actively working to address this need. The user manual provides the 25th, 50th and 75th percentile scores for each scale based on the combined sample from the first two field trials of the *DAP*. The manual cautions users that these preliminary data provide only "crude" points of comparison for research and field work with the *DAP*.

Available Supports

The *DAP* is scored by a local program administrator or evaluator (unlike their community-level surveys which are scored by Search Institute). Search Institute does not provide training for the *DAP*, so users should have experience with evaluation. Technical consultation is available from Search Institute and is negotiated on a case-by-case basis.

The user guide provides extensive information on administering, scoring and interpreting the *DAP* as well as notes on its use for research, program evaluation or clinical purposes. However, it assumes that the lead administrator has the necessary professional or academic background to interpret scores appropriately. (Search Institute suggests that masters-level training is appropriate for most applications.)

Technical Properties

This section provides information about the overall technical properties of the *DAP* and of the specific scales that map onto the skill areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these latter scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the DAP

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes

2. For what groups?

- Middle school and high school students
- Male and female youth
- Youth from different racial/ethnic groups (White, Hispanic, Asian-American, American Indian, and Multi-racial)

3. How strong is available reliability evidence?

- Substantial

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Moderate

6. What is the nature of that evidence?

- Expected differences in *DAP* scale scores for students in middle schools with contrasting levels of resources for supporting positive youth development.
- Expected associations of *DAP* scales with measures of risk behavior, thriving and grades.
- Improvements in *DAP* scale scores for youth participating in an OST program in Thailand compared to those in a random assignment control group.

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

- Does factor analysis support the scoring system for the instrument - e.g., is there support for creating separate scores for assets in each of the 8 targeted areas?
- To what extent do *DAP* scales measure their specific intended constructs - e.g., do scores on the Social Competencies scale correlate with other well-validated indices of social skills and less so with measures of abilities in other areas?
- What are the cumulative and unique contributions of *DAP* scales, when considered collectively, to the prediction of different types of youth outcomes?
- To what extent do *DAP* scales predict outcomes at later points in a youth's schooling or development?
- What is the *DAP*'s sensitivity for detecting effects of OST program participation among youth in the U.S.?

Reliability and Validity of DAP Scales Reviewed in this Guide

Scale	Number of Items	Evidence of Reliability	Evidence of Validity	Corresponding Skill Area in this Guide
Commitment to Learning	7	Substantial	Limited-to-Moderate	Initiative & Self-Direction
Social Competencies	8	Substantial	Limited-to-Moderate	Relationships & Collaboration
Positive Identity	6	Substantial	Moderate	Initiative & Self-Direction

For More Information

Jean Wachs
Search Institute
615 First Avenue NE, Suite 125
Minneapolis, MN 55413
(800) 888-7828, ext. 211
Email: jeanw@search-institute.org

Devereux Student Strengths Assessment

Overview and Purpose

The *Devereux Student Strengths Assessment (DESSA)* is a 72-item behavior rating scale designed to assess eight social-emotional competencies for children in grades K-8. The instrument is strengths-based and does not assess risk factors or maladaptive behaviors. The *DESSA* is based on a definition of social-emotional competencies, such as a child's ability to successfully interact with others in a way that demonstrates awareness of and ability to manage emotions in an age- and contextually appropriate manner. Published by the Devereux Center for Resilient Children, the *DESSA* is part of a series of strength-based assessments grounded in resilience theory that also includes the *Deveraux Early Childhood Assessment* or *DECA*.

The *DESSA-mini* is an eight-item universal screening tool that estimates a youth's overall social-emotional competence. The mini version is recommended for use in situations in which the longer form is not practical or feasible. The *DESSA-mini* does not yield individual scale scores, so programs should consider their purposes when selecting which version to use.

Content

The *DESSA* is completed by parents, teachers or program staff in child-serving settings. For each item, the rater indicates on a five-point scale (never, rarely, occasionally, frequently, very frequently) how often the student engaged in each behavior over the past four weeks. The 72 items are organized into the eight scales listed below. A Social-Emotional Composite score provides an overall assessment of the strength of a child's social-emotional competence.

The developers of the *DESSA* recommend a three-step process for interpreting scores. The first step is examining the Social-Emotional Composite as a global assessment of a child's social-emotional competencies. The second step involves reviewing the eight separate scale scores. Instructions in the manual help users convert separate scale scores into norm-based scores that can be placed into one of three categories – "strength", "typical" or "need for instruction." (For more detail, see *An Introduction to the Devereux Student Strengths Assessment*.) This step may provide useful information about the specific strengths and needs of the child. For instance, scores may suggest whether a child's strengths are primarily intrapersonal or interpersonal. Step three, individual item analysis, involves identifying strengths and needs. Overall, the preceding process may allow programs to modify both individual interventions and program-level strategies to align with children's strengths and needs.

Sample Items from *DESSA* Scales Reviewed in this Guide

During the past 4 weeks, how often did the child…

- Remember important information? (Personal Responsibility)
- Keep trying when unsuccessful? (Goal-directed Behavior)
- Get along with different kinds of people? (Social Awareness)
- Give an opinion when asked? (Self-awareness)
- Wait for her/his turn? (Self-management)
- Compliment or congratulate somebody? (Relationship Skills)
- Learn from experience? (Decision Making)

The *DESSA Record Form* displays scores in two graphic formats: the Individual Student Profile, which conveys strengths and needs compared to national norms, and the Classroom/Program Profile, which depicts social-emotional functioning against national norms of all participants in a given classroom or program group.

The *DESSA* includes the following scales:

- Self-Awareness*
- Social-Awareness*
- Self-Management*
- Goal-Directed Behavior*
- Relationship Skills*
- Personal Responsibility*
- Decision Making*
- Optimistic Thinking

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

User Considerations

This section discusses the *DESSA* in terms of several important user considerations, including accessibility, ease of use, availability of normative data and supports available to programs.

Accessibility

The *DESSA* may be purchased through Kaplan (www.k5kaplan.com). A standard kit costs $115.95 and includes a user manual, a norms reference card and 25 hand scoring forms. Additional packages of 25 forms may be purchased for $39.95 each. An online *DESSA* scoring assistant costs $32.25 for 25 online forms. The *DESSA* is available in Spanish.

Ease of Use

The *DESSA* is a filled out by an adult – a teacher, program staff member or parent – for each child being assessed. It takes approximately 10-15 minutes per child. Programs should consider their time and human resource constraints for completing the forms, as the *DESSA* is not a self-report tool.

Availability of Norms

Normative data that characterize what is usual within a defined population can help programs better understand the populations they serve and the effects of their programs. Normative data are available for each scale of the *DESSA* and the Social-Emotional Composite, based on a standardization sample consisting of nearly 2,500 children. The sample is reported to closely approximate the K-8 population of the U.S. with respect to age, gender, geographic region of residence, race, ethnicity and socioeconomic status based on data published in 2008 by the U.S. Census Bureau. Norm reference cards are available for purchase and are included in the *DESSA* kit.

Available Supports

The user manual offers detailed instructions for users. Programs seeking more information prior to purchase may read an introduction to the *DESSA*. Fee-based in-service training is available but not required. Free video and audio training presentations are available at www.centerforresilientchildren.org.

Technical Properties

This section provides information about the overall technical properties of the *DESSA* and of specific scales that map onto the skill areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these seven scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the DESSA

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes

2. For what groups?

- Elementary school students (grades K-8 collectively)

3. How strong is available reliability evidence?

- Moderate

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Limited-to-Moderate

6. What is the nature of that evidence?

- For selected scales on the *DESSA*, relatively strong correlations with scales on other instruments that assess abilities or behaviors in similar areas.
- Expected associations of *DESSA* scale and composite scores with teacher ratings of youth emotional, behavioral, and school functioning on other established measures (Criterion validity).

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

- Does factor analysis support the scoring system for the instrument - e.g., is there support for creating separate scores for skills in each of the 8 targeted areas?
- To what extent do *DESSA* scales measure their specific intended constructs - e.g., does the Decision Making scale correlate with other established measures of skills in this area and less so with measures of other types of skills?
- What are the cumulative and unique contributions of *DESSA* scales, when considered collectively, to the prediction of different types of youth outcomes?
- What is the instrument's sensitivity for detecting expected effects of OST program participation?
- Do ratings by OST program staff on the *DESSA* exhibit evidence of validity?

** The scope of this assessment of the reliability and validity of the DESSA does not include ratings on the instrument that are provided by the child's parent.*

Reliability and Validity of Specific *DESSA* Scales				
DESSA Scale	**Number of Items**	**Evidence of Reliability**	**Evidence of Validity**	**Corresponding Skill Area in this Guide**
Self-Awareness	7	Moderate	Limited-to-Moderate	Initiative & Self-Direction
Social Awareness	9	Moderate	Limited-to-Moderate	Relationships & Collaboration
Self-Management	11	Moderate	Moderate	Relationships & Collaboration
Goal Directed Behavior	10	Moderate	Limited-to-Moderate	Initiative & Self-Direction
Relationship Skills	10	Moderate	Moderate	Relationships & Collaboration
Personal Responsibility	10	Moderate	Limited-to-Moderate	Initiative & Self-Direction
Social Awareness	9	Moderate	Limited-to-Moderate	Relationships & Collaboration
Decision Making	8	Moderate	Limited-to-Moderate	Critical Thinking & Decision Making

For More Information

Paul A. LeBuffe, M.A.
Devereux Center for Resilient Children
444 Devereux Drive, P.O. Box 638
Villanova, PA 19085
(610) 542-3090
(610) 542-3132 (f)
plebuffe@Devereux.org

San Francisco Beacons Youth Survey

Overview and Purpose

The *San Francisco Beacons Youth Survey (Beacons Youth Survey)* was developed by Public/Private Ventures (P/PV) as part of an effort to evaluate the first five Beacons centers that opened in San Francisco between the 1996 and 1998. This self-report survey is designed to assess how middle school youth spend their out-of-school time (e.g., time in challenging activities) and to document developmental outcomes related to their well-being (such as self-efficacy).

The *Beacons Youth Survey* is designed for OST programs and was created for research purposes. As such it has not been widely distributed beyond the San Francisco effort. P/PV has also developed a staff survey and other tools that programs can use to link information about activity quality, participation and youth experiences.

Content

The different scales on the *Beacons Youth Survey* include items with a range of different response formats. The most common format asks youth to respond using a four-point scale: strongly agree, somewhat agree, somewhat disagree, strongly disagree.

The survey consists of 10 scales plus questions about basic demographic information. The scales are:

- School Effort*
- Self-Efficacy*
- Positive Reaction to Social Challenge*
- Passive Reaction to Social Challenge
- Leadership*
- Non-Familial Support
- Peer Support
- Time Spent in Challenging Learning Activities*
- Adult Support at the Beacons
- Variety of Interesting Activities offered at the Beacons

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

Sample Items from *San Francisco Beacons Youth Survey* Scales Reviewed in this Guide

I pay attention in class
(School Effort)

I can depend on myself
(Self-Efficacy)

When I have a problem or argument with another student, I think about it afterward and try to figure out what went wrong
(Positive Reaction to Social Challenge)

In the last year, how often have you helped plan activities or events for a group, team or club?
(Leadership)

Art, music, dance or drama class or lesson
(Time Spent in Challenging Learning Activities)

User Considerations

This section discusses the *Beacons Youth Survey* in terms of several important user considerations including accessibility, ease of use, availability of normative data and supports available to programs.

Accessibility

The *Beacons Youth Survey* is available free of charge, though it was designed originally for research purposes and has not been adapted or packaged specifically for practitioner use. If non-Beacon programs use the tool, they can replace references to "Beacons" (in the Beacons Experience scales) with the name of their program. More information about this survey is outlined in a program evaluation report that describes its use in the Beacons programs.

Ease of Use

A typical youth can complete the *Beacons Youth Survey* in about 35 minutes. The survey is designed to be read aloud to youth in groups and filled out using paper and pencil. The survey is intended to be used in its entirety, although individual scales can be used alone as well.

Availability of Norms

Normative data designed to facilitate comparison of youth in a given program to a larger population are not available.

Available Supports

P/PV does not offer training to survey users. Programs interested in using the tool can contact the developer for limited guidance on administration. However, programs will have to collect and analyze their own data and should seek out an experienced local evaluator for assistance if necessary.

P/PV has developed a companion *Youth Feedback Form* designed to assess the quality of youths' program experiences. This survey may be used in tandem with the youth survey for programs interested in gathering additional data to guide program improvement.

Technical Properties

This section provides information about the overall technical properties of the *San Francisco Beacons Youth Survey* and of specific scales that map onto the skill areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these latter scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of San Francisco Beacons Youth Survey

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes

2. For what groups?

- Primarily for middle school-age youth

3. How strong is available reliability evidence?

- Limited-to-Moderate

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Moderate

6. What is the nature of that evidence?

- In path modeling analyses, several of the scales were linked to improvements in school grades.
- For some scales, expected increases over time in association with OST program participation.

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

- To what extent do scales on the *Beacons Youth Survey* measure their specific intended constructs - e.g., does the Self-Efficacy scale correlate with other established indices of this construct and less so with measures of youth attitudes or skills in other areas?
- To what degree do *Beacons Youth Survey* scales contribute to the prediction of youth outcomes in non-academic domains?
- What is the sensitivity of scales on the *Beacons Youth Survey* for detecting effects of OST program participation?

**This summary does not include the scales on the Beacons Youth Survey that ask youth to report on their after-school program experiences (i.e., Adult Support at the Beacons and Variety of Interesting Activities offered at the Beacons).*

Reliability and Validity of Specific *San Francisco Beacons Youth Survey* Scales

Scale	Number of Items	Evidence of Reliability	Evidence of Validity	Corresponding Skill Area in this Guide
School Effort	4	Limited	Moderate	Initiative & Self-Direction
Self-Efficacy	8	None	Limited-to-Moderate	Initiative & Self-Direction
Positive Reaction to Social Challenge	6	Limited	Moderate	Relationships & Collaboration
Leadership	11	None	None-to-Limited	Initiative & Self-Direction
Time Spent in Challenging Learning Activity	8	None	Limited	Initiative & Self-Direction

For More Information

Amy Arbreton
Public/Private Ventures
Lake Merritt Plaza
1999 Harrison Street
Oakland, CA 94612
(510) 273-4600
AArbreton@PPV.org

Survey of After-School Youth Outcomes

Overview and Purpose

The *Survey of After-School Youth Outcomes (SAYO)* was developed by the National Institute on Out-of-School Time (NIOST) in 2003, in partnership with the Massachusetts Department of Elementary and Secondary Education for the 21st Century Community Learning Centers program. Updated in 2007, the *SAYO* is designed to collect data about youth from teachers, OST program staff and youth about intermediary youth outcomes that link to long-term healthy development and educational success.

The staff and teacher surveys are called the *SAYO-S* and *SAYO-T*. There are two versions of the *SAYO-Y*, for grades 4-8 and 9-12. The *SAYO* is part of the *Afterschool Program Assessment System (APAS)*, which includes an observational measure of program quality.

Content

The *SAYO-S & -T* are based on a menu approach and programs are encouraged to collect data on outcomes that are most aligned with their goals. The *SAYO-Y* includes scales assessing youths' experiences in an OST program as well as outcomes in the areas of sense of competence and future planning and expectations.

The *SAYO-S & -T* each contain more than 30 questions organized into eight and nine scales respectively. The items use a five-point response scale: never, rarely, sometimes, usually, always. *SAYO-Y* scales target areas considered by the developers to be best measured by asking youth directly. The two versions of the *SAYO-Y* each contain more than 80 questions divided across 18 scales. Students report on a range of their own perceptions, beliefs and attitudes a four-point response scale: no, mostly no, mostly yes, yes.

SAYO-S & -T scales include:

- Behavior in the Program (*SAYO-S* only)
- Behavior in the Classroom* (*SAYO-T* only)
- Initiative*
- Engagement in Learning
- Relations with Adults*
- Relations with Peers*
- Problem Solving Skills*
- Communication Skills*
- Homework
- Academic Performance (*SAYO-T* only)

Sample Items from *SAYO* Scales Reviewed in this Guide

Seeks appropriate assistance and support from teacher or other adults in resolving problems (Relations with Adults, *SAYO-T*)

Initiates interactions with adults (Relations with Adults, *SAYO-S*)

Shows consideration for peers (Relations with Peers, *SAYO-T/SAYO-S*)

Demonstrates active listening skills (e.g., is able to summarize key points of speaker) (Communication Skills, *SAYO-T/SAYO-S*)

Is able to regain control of behavior when given a warning (Behavior in the Classroom, *SAYO-T*)

Sets goals for self (Initiative, *SAYO-T/SAYO-S*)

When encounters difficulty, is able to identify and describe the problem (Problem Solving Skills, *SAYO-T/SAYO-S*)

It's easy for me to join a new group of teens (Sense of Competence Socially, *SAYO-Y*)

I set goals for myself. For instance, things I want to learn or get better at. (Future Planning – My Actions, *SAYO-Y*)

SAYO-Y scales cluster into three broad areas:

Program Experiences
- Engagement and Enjoyment
- Choice and Autonomy
- Challenge
- Perceptions of the Social Environment
- Supportive Relationships with Staff Members
- Responsibility and Leadership

Future Planning and Expectations
- Future Planning – My Actions*
- Expectations
- Aspirations and College Planning

Sense of Competence
- Sense of Competence in Reading
- Sense of Competence in Writing
- Sense of Competence in Math
- Sense of Competence in Science
- Sense of Competence as a Learner
- Sense of Competence Socially*

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

.

User Considerations

This section discusses the *SAYO* in terms of several important user considerations including accessibility, ease of use, availability of normative data and supports available to users.

Accessibility

Programs receive a one-year license to use any or all of the *SAYO* surveys after participating in required training. An online option is available for $250; more information is provided below under "available supports." The *SAYO-Y* is administered online only. The staff and teacher versions can be administered online or using paper/pencil.

Ease of Use

The *SAYO* surveys contain more questions than are recommended for a single administration. NIOST recommends that programs customize the survey by selecting scales that best fit their goals. In addition to selecting which scales to use, programs may choose packages that include either the youth, teacher or staff versions, or a combination.

For the *SAYO-Y*, programs are encouraged to select scales that sum to no more than 50 questions total. Programs are encouraged to choose only three outcome scales when using either the staff or teacher surveys.

Availability of Norms

Normative data designed to facilitate comparison of youth in a given program to a larger population are not currently available.

Available Supports

NIOST has a range of supports available for the full *APAS* system, which includes the *SAYO* as well as the previously noted observational program quality assessment tool, called the *Afterschool Program Practices Tool (APT)*. Though the tools are designed to work in tandem to help improve program quality and outcomes, the *SAYO* can be used as an outcome measure apart from the *APT*.

To use the *SAYO*, programs must participate in a NIOST training (two staff per site are recommended) or take an online tutorial. The online tutorial costs $250 and includes one year of online access to tools. NIOST does not require that *SAYO* users have prior research or evaluation experience. Additional information on training can be found at the NIOST website.

Technical Properties

This section provides information about the overall technical properties of the *SAYO* and of specific scales that map onto the areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these latter scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the SAYO*

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes.

2. For what groups?

- Elementary/middle and high school students
- Male and female youth
- Youth from different racial/ethnic groups (White, Black, Hispanic and Asian-American)

3. How strong is available reliability evidence?

- Substantial.

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes.

5. How strong is available validity evidence?

- Moderate-to-Substantial.

6. What is the nature of that evidence?

- Convergence between ratings from teachers and OST program staff on corresponding scales of the *SAYO-T* and *SAYO-S*.
- Expected associations of scales with teacher ratings of the quality of the youth's school work and academic performance.
- Expected associations of scales with youth reports of their OST program experiences and of their academic and personal/social gains associated with program participation.
- Expected patterns of differential improvement for scale scores on the *SAYO-S* in association with participation in OST programs of varying quality.
- Support for *SAYO* scales as intervening variables in pathways linking youth reports of their OST experiences to teacher reports of their academic performance.

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

- Does factor analysis support the scoring system for the instrument – e.g., is there support for creating separate scores for each of the areas that are assessed on each informant version of the *SAYO*?
- To what extent do *SAYO* scales measure their specific intended constructs – e.g., does the Communications Scale correlate with other established measures of skills in this area and less so with measures of other types of skills?
- To what degree do scales on the instrument predict measures of youth outcomes from outside of the *SAYO* assessment system?
- To what extent do *SAYO* scales predict outcomes at later points in a youth's schooling or development?
- What is the sensitivity of scales on the *SAYO* for detecting effects of OST program participation when utilizing a quasi-experimental or randomized control evaluation design?

** This summary does not include scales on the* SAYO *that typically would be viewed as indices of more distal youth outcomes such as the scale on the* SAYO-T *in which teachers rate the youth's academic competence, or those scales on the* SAYO-Y *that are focused on the youth's program experiences.*

Reliability and Validity Evidence for *SAYO* Scales Reviewed in this Guide

Scale	Number of Items	Evidence of Reliability	Evidence of Validity	Corresponding Skill Area in this Guide
Behavior in the Classroom *(SAYO-T)*	4	Moderate-to-Substantial	Moderate-to-Substantial	Initiative & Self-Direction
Initiative *(SAYO-S & -T)*	5	Substantial	Moderate-to-Substantial	Initiative & Self-Direction
Relations with Adults *(SAYO-S & -T)*	4 or 5	Substantial	Moderate-to-Substantial	Relationships & Collaboration
Relations with Peers *(SAYO-S & -T)*	3 or 4	Substantial	Moderate-to-Substantial	Relationships & Collaboration
Problem-Solving Skills *(SAYO-S & -T)*	3 or 5	Substantial	Moderate-to-Substantial	Critical Thinking & Decision Making
Communication Skills *(SAYO-S & -T)*	4 or 5	Substantial	Moderate-to-Substantial	Communication
Future Planning – My Actions *(SAYO-Y)*	4	Substantial	Moderate-to-Substantial	Initiative & Self-Direction
Sense of Competence Socially *(SAYO-Y)*	4	Substantial	Moderate-to-Substantial	Relationships & Collaboration

For More Information

Wendy Surr
National Institute on Out-of-School Time
Wellesley Centers for Women
Wellesley College
106 Central Street
Wellesley, MA 02481
(781) 283-2443
Email: wsurr@wellesley.edu

Social Skills Improvement System Rating Scales

Overview and Purpose

The *Social Skills Improvement System (SSIS)* is a multi-tiered assessment and intervention system aimed at supporting youths' social skills. The suite of tools focuses on skills that enable academic and social success for youth ages 3-18. The *SSIS Rating Scales* replace an earlier instrument called the *Social Skills Rating System*.

The *SSIS* includes rating scales, a performance screening guide, an intervention guide and a class-wide intervention program. The rating scales, which are the focus of our review, utilize a multi-rater approach in which students, teachers and parents provide parallel assessment information for each youth being assessed.

Content

The *SSIS Rating Scales* capture student, teacher and parent reports on the "frequency and perceived importance of positive behaviors" as well as information on problem behaviors that may interfere with a student's ability to demonstrate prosocial skills. Teachers also provide ratings of the student's academic competence.

The Teacher and Parent Forms allow for rating youth as young as age 3 up through age 18. There are two self-report Student Forms, for ages 8-12 and 13-18. The number of items averages about 80 items per form, but varies somewhat based on the form and age of the youth.

The Teacher and Parent Forms ask raters to indicate the frequency of behaviors demonstrated by youth on a four-point scale: never, seldom, often, almost always. Youth are asked how true various statements are for them: not true, a little true, a lot true, very true. Teachers, parents and older students (13-18) are also asked to rate the importance of each social skills behavior to the student's development on a three-point scale: not important, important, critical.

Administrators can use a summary sheet for each form to calculate an overall set of ratings for individual youth. For each domain an individual youth's score is categorized as well-below average, below average, average, above average, or well-above average based on a comparison to normative data. The user manual outlines procedures and examples for interpreting reports and reporting when there are multiple raters.

Sample Items from *SSIS* Subscales Reviewed in this Guide

Teacher Form

Speaks in appropriate tone of voice (Communication)

Stands up for others who are treated unfairly (Assertion)

Tries to comfort others (Empathy)

Participates in games or group activities (Engagement)

Stays calm when teased (Self-Control)

Student Form

I am polite when I speak to others (Communication)

I stand up for others who are not treated well (Assertion)

I try to make others feel better (Empathy)

I smile or wave at people when I see them (Engagement)

I stay calm when I am teased (Self-Control)

The *SSIS* includes three scales with corresponding subscales for two of the scales:

- Social Skills
  - Communication*
  - Cooperation
  - Assertion*
  - Responsibility
  - Empathy*
  - Engagement*
  - Self-Control*
- Competing Problem Behaviors
  - Externalizing
  - Bullying
  - Hyperactivity/Inattention
  - Internalizing
  - Autism Spectrum (teacher and parent report only)
- Academic Competence (teacher report only)

** These subscales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

User Considerations

This section discusses the *SSIS Rating Scales* in terms of several important considerations including accessibility, ease of use, availability of normative data and supports available to programs.

Accessibility

The *SSIS Rating Scales* are distributed through Pearson. The parent and student versions are available in English and Spanish. Users may purchase either hand-scored or computer-scored starter kits. The hand-scored starter kit costs $248.25 and includes a user manual and three packages of 25 student, teacher and parent forms. The computer-scored starter kit costs $517.35 and includes the manual, a package of each set of forms and scoring software. Packets of 25 additional forms are available and cost $43.05 (hand-scored) and $53.60 (computer scored).

Ease of Use

Each form takes 10 to 25 minutes to complete. No special training is required to administer the scale, and procedures for scoring are outlined in the user guide.

Availability of Norms

The *SSIS* Rating Scale has been tested on a normative sample of 4,700 youth ages 3-18. In addition, 385 teachers and 2,800 parents provided ratings. Sampling was conducted on a national standardization sample aligned with the demographic data published in 2006 by the U.S. Census Bureau. The three forms have normative scores by age group and gender. Information about using norms is included in kits.

Available Supports

The user guide includes information on administering, scoring and interpreting results. The manual suggests that interpretation of scores and reports should be done by a professional familiar with test construction and interpretation (an evaluator, for example), as no additional training is provided.

These scales are part of a family of assessment and intervention tools, including a universal screening tool and social skills intervention programs. These other tools may be purchased to use in tandem with the rating scales. Programs purchasing the computer-scored kit may link directly to specific interventions based on scores obtained for individual youth.

Technical Properties

This section provides information about the overall technical properties of the *SSIS Rating Scales* and of specific scales that map onto the skill areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these latter scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the SSIS Rating Scales*

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes

2. For what groups?

- Male and female youth ages 12 and under, and ages 13-18

3. How strong is available reliability evidence?

- Moderate-to-Substantial

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Moderate

6. What is the nature of that evidence?

- Ratings for *SSIS* scales and subscales on the teacher and youth forms typically have exhibited convergence with ratings of other informants (youth and parent informants for teacher ratings and teacher and parent informants for youth ratings) for the corresponding scale or subscale.
- For the most part, *SSIS* scales and subscales have exhibited expected associations with concurrent measures of youth emotional, behavioral and academic functioning.

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

- Does factor analysis support the scoring system for the instrument – e.g., is there support for creating separate scores for social skills in each of the targeted areas?
- Do teacher- and youth-report subscales on the *SSIS* measure their specific intended constructs – for example, does the Empathy subscale correlate with other well-validated indices of skills in this area and less so with measures of other types of skills?
- What are the cumulative and unique contributions of *SSIS* scales and subscales, when considered collectively, to the prediction of different types of youth outcomes?
- To what extent do *SSIS* scales and subscales predict outcomes assessed at later points in a youth's schooling or development?
- What is the sensitivity of the *SSIS Rating Scales* for detecting expected effects of OST program participation?

** This summary encompasses only the Social Skills scale and associated subscales of the* SSIS Rating Scales. *The Problem Behaviors scale and associated subscales and the Academic Competence scale are not included, as these typically would be viewed as indices of broader youth outcomes that are not the focus of this guide. Furthermore, in keeping with the focus of this guide on tools for use by OST programs, the summary pertains only to the student and teacher versions of the scale (i.e., does not include the parent version).*

Reliability and Validity Evidence for *SSIS* Scales Reviewed in this Guide

Scale	Number of Items	Evidence of Reliability	Evidence of Validity	Corresponding Skill Area in this Guide
Communication (Teacher)	7	Moderate-to-Substantial	Moderate	Communication
Communication (Student)	6	Moderate-to-Substantial	Moderate	Communication
Assertion (Teacher)	7	Moderate	Moderate	Relationships & Collaboration
Assertion (Student)	7	Moderate-to-Substantial	Moderate	Relationships & Collaboration
Empathy (Teacher)	6	Moderate	Moderate	Relationships & Collaboration
Empathy (Student)	6	Moderate-to-Substantial	Moderate	Relationships & Collaboration
Engagement (Teacher)	6	Moderate-to-Substantial	Moderate	Relationships & Collaboration
Engagement (Student)	6	Moderate-to-Substantial	Moderate-to-Substantial	Relationships & Collaboration
Self-Control (Teacher)	7	Moderate-to-Substantial	Moderate	Relationships & Collaboration
Self-Control (Student)	6	Moderate-to-Substantial	Moderate	Relationships & Collaboration

For More Information

Rob Altmann
Pearson
5601 Green Valley Drive
Bloomington, MN 55437
(952) 681-3268
Rob.Altmann@pearson.com

Youth Outcomes Battery

Overview and Purpose

The American Camping Association (ACA) *Youth Outcomes Battery* is a series of surveys that measure 11 youth outcome areas. Developed primarily for camp settings, the surveys are also intended to be applicable to other settings focused on supporting youth development. ACA encourages using the *Youth Outcomes Battery* to evaluate program goals and in conjunction with quality improvement efforts.

Content

The *Youth Outcomes Battery* includes three survey tools: a *Camper Learning Scale* for 6- to 9-year-olds and Basic and Detailed versions of a *Camp Youth Outcomes Scales* for 10- to 17-year-olds. Users can administer different combinations of scales from these tools depending on their focal outcomes.

The *Camper Learning Scale* includes 14 questions that ask youth about how much they learned in different areas during their camp experience. The Basic version of the *Camp Youth Outcomes Scales* is recommended for youth ages 10-13. It includes approximately 65 questions that ask youth about how much their camp experienced changed their levels of skills in different areas (see list below). The Detailed version of the *Camp Youth Outcomes Scales* is recommended for older youth (13-17). The questions on this version are parallel in content to those on the Basic version, but each question has two parts so as to assess both current "status" and "change." The first part asks youth how true the statement is of them ("status") using a six-point response scale: false, somewhat false, a little false, a little true, somewhat true, true. The second part asks youth to report how much or less true it is of them now compared to before they came to camp ("change"), using another six-point response scale: a lot less, somewhat less, a little less, a little more, somewhat more, a lot more.

Sample Items from *Youth Outcomes Battery* Scales Reviewed in this Guide

I am good at trusting my friends
(Friendship Skills)

I am good at taking care of myself
(Independence)

I know I can get along with other people in a small group
(Teamwork)

I want to learn more about new ideas
(Interest in Exploration)

I don't blame others for my mistakes
(Responsibility)

When I have a problem, I make good choices about what to do
(Problem-Solving Confidence)

Finally, there is a Camp Connectedness scale that can be administered with both the Basic and Detailed versions of the *Camp Youth Outcome Scales*. This scale measures the camper's personal relationship to camp in areas such as belonging, youth voice and staff support. For purposes of this guide, the only scales reviewed are the "status" scales from Detailed version of the *Camp Youth Outcome Scales.*

Responses to items are scored from 1-6 in ascending order of response choice. Scale scores are then calculated by summing the scores for each item on a given scale. The results can be used to describe perceived outcomes of youth and can be broken down by other variables, such as age of youth or program type.

The *Youth Outcomes Battery* includes the following scales:

- Friendship Skills*
- Independence*
- Teamwork*
- Family Citizenship
- Perceived Competence
- Interest in Exploration*
- Responsibility*
- Problem-Solving Confidence*
- Affinity for Nature
- Spiritual Well-being
- Camp Connectedness

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

User Considerations

This section discusses the *Youth Outcomes Battery* in terms of several important considerations including accessibility, ease of use, availability of normative data and supports available to users.

Accessibility

The full set of tools can be purchased online by ACA members for $40 and by non-members for $120. Organizations may also purchase individual scales for $5 (members) or $15 (non-member). Once purchased, programs can make as many copies as they need. Users also have access to an online Excel-based data analysis template on the ACA website.

Ease of Use

The survey is available in a paper/pencil format. The basic version for older youth takes five to 20 minutes, depending on the number of scales administered. The detailed version requires more time because each question has two parts. Non-camp programs will need to adapt camp-specific language to fit their program context.

Availability of Norms

Normative data that characterize what is usual within a defined population can help programs better understand the populations they serve and the effects of their programs. ACA recently began collecting normative data on the Basic version of the *Youth Outcomes Battery* in which youth report retrospectively on the extent to which their skills have changed in different areas as a result of their camp experience. These data are intended to allow individual camps to compare their scores with representative scores from typical ACA camps. (The data offer limited comparison value for non-residential camp programs because 75 percent were collected in residential camps.) Additional work is underway and details related to gender, age, race/ethnicity and day/resident programming are forthcoming. Guidance on how to use norms for comparison purposes is available at www.acacamps.org/research/enhance/youth-outcomes-resources/norms.

Available Supports

Although ACA does not provide training to programs outside of their membership, it has developed written guidelines for the administration and scoring of the instruments and data analysis. The user guide outlines differences between survey versions, tips for administering and scoring and scripts for staff to follow when administering the survey. As noted above, users also have access to an Excel template to help with data analysis. See the ACA website for additional information.

Technical Properties

This section provides information about the overall technical properties of the *Youth Outcomes Battery* and of the specific scales that map onto the areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these latter scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the Youth Outcomes Battery

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes

2. For what groups?

- Reliability findings have not been reported for specific groups of youth

3. How strong is available reliability evidence?

- Limited

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Limited

6. What is the nature of that evidence?

- Expected associations of scale scores with youth ratings of their change in the corresponding areas since coming to camp.

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

- Does factor analysis support the scoring system for the instrument – eg., is there support for creating separate scores for each of the targeted areas?
- Do scales measure their specific intended constructs – e.g., do scores on the Friendship Skills scale correlate with other well-validated measures of social competence and less so with measures that target skills in other areas?
- To what extent are *YOB* scales useful in predicting other important youth outcomes?
- What is the *Youth Outcomes Battery* sensitivity for detecting effects of OST program participation?

** This summary is limited to the status format scales on the Detailed version of the* Camp Youth Outcome Scales.

Reliability and Validity of *Youth Outcomes Battery* Scales Reviewed in this Guide				
Scale	**Number of Items**	**Evidence of Reliability**	**Evidence of Validity**	**Corresponding Skill Area in this Guide**
Friendship Skills	13	Limited	Limited	Relationships & Collaboration
Independence	8	Limited	Limited	Initiative & Self-Direction
Teamwork	8	Limited	Limited	Relationships & Collaboration
Interest in Exploration	8	Limited	Limited	Initiative & Self-Direction
Responsibility	6	Limited	None-to-Limited	Initiative & Self-Direction
Problem-Solving Confidence	8	Limited	Limited	Critical Thinking & Decision Making

For More Information

M. Deborah Bialeschki, Ph.D.
Director of Research
American Camp Association
5000 State Road 67 North
Martinsville, IN 46151
(765) 349-3318
dbialeschki@acacamps.org
www.acacamps.org/research

Youth Outcome Measures Online Toolbox

Overview and Purpose

The *Youth Outcome Measures Online Toolbox (Online Toolbox)* is a battery of measures that assesses positive behavior change and skill development in youth. Based on research about out-of-school time participation, the measures have been adapted and organized into an online platform by researchers Deborah Vandell, Kim Pierce, Pilar O'Cadiz, Valerie Hall, Andrea Karsh and Teresa Westover. The *Online Toolbox* contains a set of measures to be completed by program staff, school day teachers, and elementary and middle school students.

Content

Teacher and staff surveys provide parallel perceptions of individual youth and when administered on multiple occasions over time, are designed to yield a comprehensive picture of behavior change and skill development. The teacher and staff surveys each contain 44 questions that ask these adults to rate youth in terms of specific behaviors (see sample items). Most questions use a five-point response scale: very poor, somewhat poor, average, good, very good. The youth survey contains 30 questions that ask young people how true a given statement is about them: not at all true, a little true, mostly true, really true. The battery is intended to be used in its entirety, although individual scales can stand alone.

The staff and teacher surveys include the following scales:

- Social Skills*
- Prosocial Behavior with Peers*
- Aggressive Behavior with Peers
- Work Habits*
- Task Persistence*
- Academic Performance (teacher version only)

The youth survey includes these scales:

- Social Competencies*
- Misconduct
- Work Habits*
- Reading/English Efficacy
- Math Efficacy

** These scales each map onto one of the skill areas that are the focus of this guide. The Technical Properties section below summarizes our ratings of the reliability and validity evidence for these scales.*

Sample Items from *Youth Outcome Measures Online Toolbox* Scales Reviewed in this Guide

Understands others' feelings
(Social Skills – Teacher/Staff)

Generates many solutions to interpersonal problems
(Prosocial Behavior with Peers – Teacher/Staff)

This student uses time wisely
(Work Habits – Teacher/Staff)

This student gives up on things before finishing them
(Task Persistence – Teacher/Staff)

I can tell a funny story to a group of kids
(Social Competencies – Youth)

I work well by myself
(Works Habits – Youth)

User Considerations

This section discusses the *Youth Outcome Measures Online Toolbox* in terms of several important user considerations, such as accessibility, ease of use, availability of normative data and supports available to programs.

Accessibility

Information and resource materials about the *Online Toolbox* are available at http://afterschooloutcomes.org/.

Programs interested in using the measure independently are free to do so. To receive a list of the survey items, contact the tool developers via the website or by e-mailing afterschool@uci.edu. This free list of scales and survey items is not in survey format; it is meant for interested parties to view and use independently. Programs interested in using the *online toolbox* portal need to enter into a service agreement with University of California at Irvine. A one-year service agreement includes 1) access to the online surveys; 2) technical assistance in administering the surveys; and 3) analysis and reporting of data. Fees vary based on the number of sites, number of students per site and level of analyses.

Ease of Use

The surveys in the *Online Toolbox* can be administered online or using paper/pencil. The developers report that most youth can complete the battery in about 10 minutes and that most teachers and program staff can complete ratings on one student in five to eight minutes.

Availability of Norms

Tables with normative data designed to facilitate comparison of youth in a given program to a larger population are not currently available.

Available Supports

Minimal training, i.e., self-training by reading instructions on the project website, is necessary to administer these measures. Step-by-step instructions and additional resource materials are available at http://afterschooloutcomes.org at no cost. Programs can enter into a fee-based service agreement with the research team for access to the *Online Toolbox*, ongoing support via telephone and e-mail, and data analysis.

Further information about the *Online Toolbox* is included in two reports (Vandell et al., 2010).[ix]

Technical Properties

This section provides information about the overall technical properties of the *Online Toolbox* and of specific scales that map onto the skill areas that are the focus of this guide. The Technical Appendix provides detailed analysis of reliability and validity evidence for these latter scales as well as a description of the process used to arrive at ratings.

Reliability and Validity of the Youth Outcome Measures Online Toolbox

1. Is there evidence that the scales on the instrument generate consistent responses, that is, are reliable?

- Yes.

2. For what groups?

- Elementary and middle school students
- Male and female youth
- English Language Learner youth
- Youth from different racial/ethnic groups (White, Black, Hispanic, and Asian-American)

3. How strong is available reliability evidence?

- Substantial

4. Is there evidence that the scales on the instrument measure what they intend to measure, that is, are valid?

- Yes

5. How strong is available validity evidence?

- Moderate

6. What is the nature of that evidence?

- Convergence of ratings from teachers and OST program staff for the same *Online Toolbox* scales.
- Associations of selected *Online Toolbox* scales with established measures of the same or similar constructs.
- Associations of *Online Toolbox* scales with relevant criterion or outcome measures such as academic achievement test scores.
- Expected patterns of improvement in *Online Toolbox* scale scores in association with OST program participation.

7. What are some of the questions that it would be useful for scholars to address as they continue to work with this instrument?

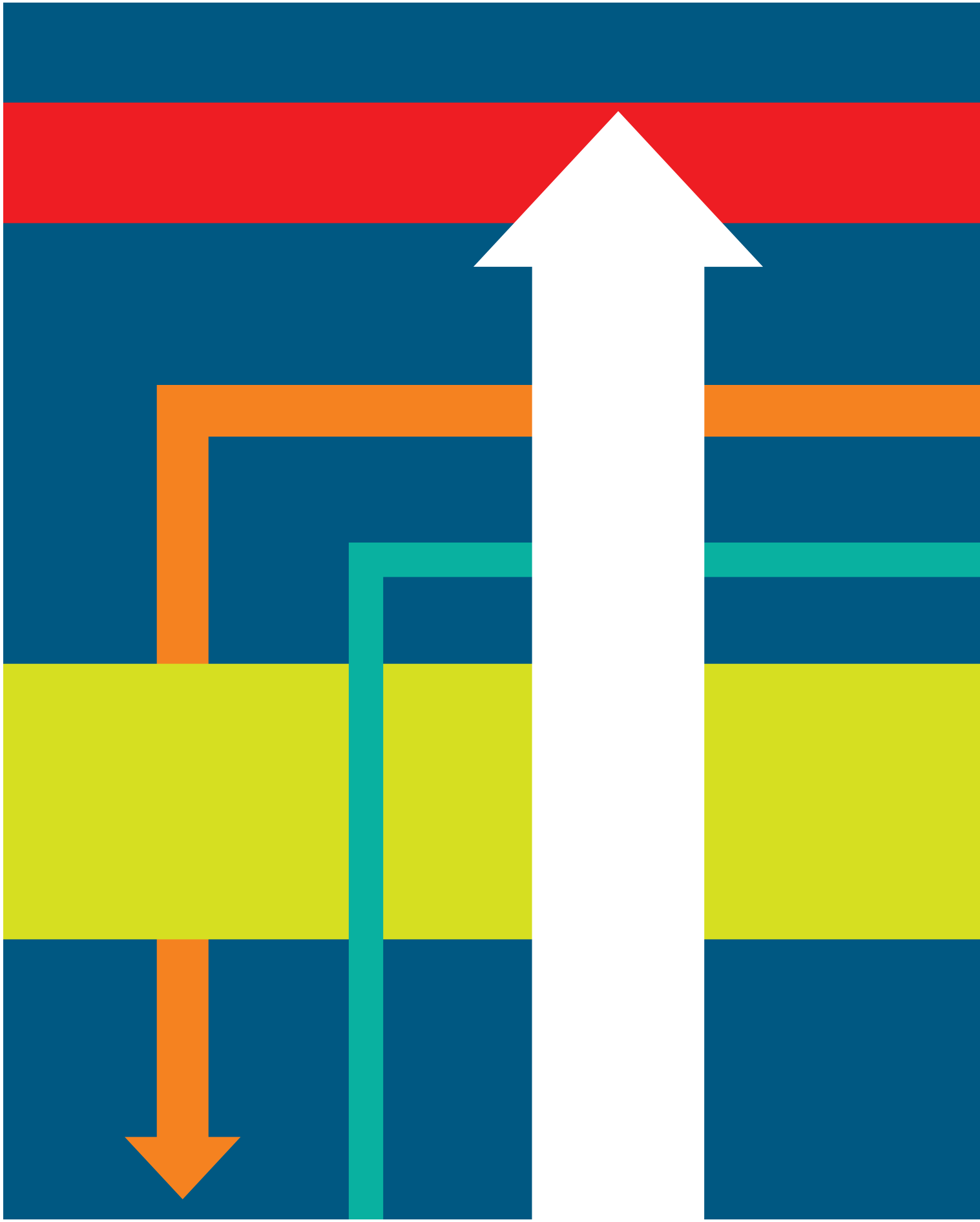
- To what extent do *Online Toolbox* scales measure their specific intended constructs – e.g., does the Social Skills scale measure a distinct construct from the Prosocial Behavior scale, with which it has demonstrated a high level of association?
- What are the cumulative and unique contributions of *Online Toolbox* scales, when considered collectively, to the prediction of different types of youth outcomes?
- To what extent do the *Online Toolbox* scales predict outcomes assessed at later points in a youth's schooling or development?
- What is the sensitivity of scales in the *Online Toolbox* for detecting effects of OST program participation when utilizing a randomized control evaluation design?

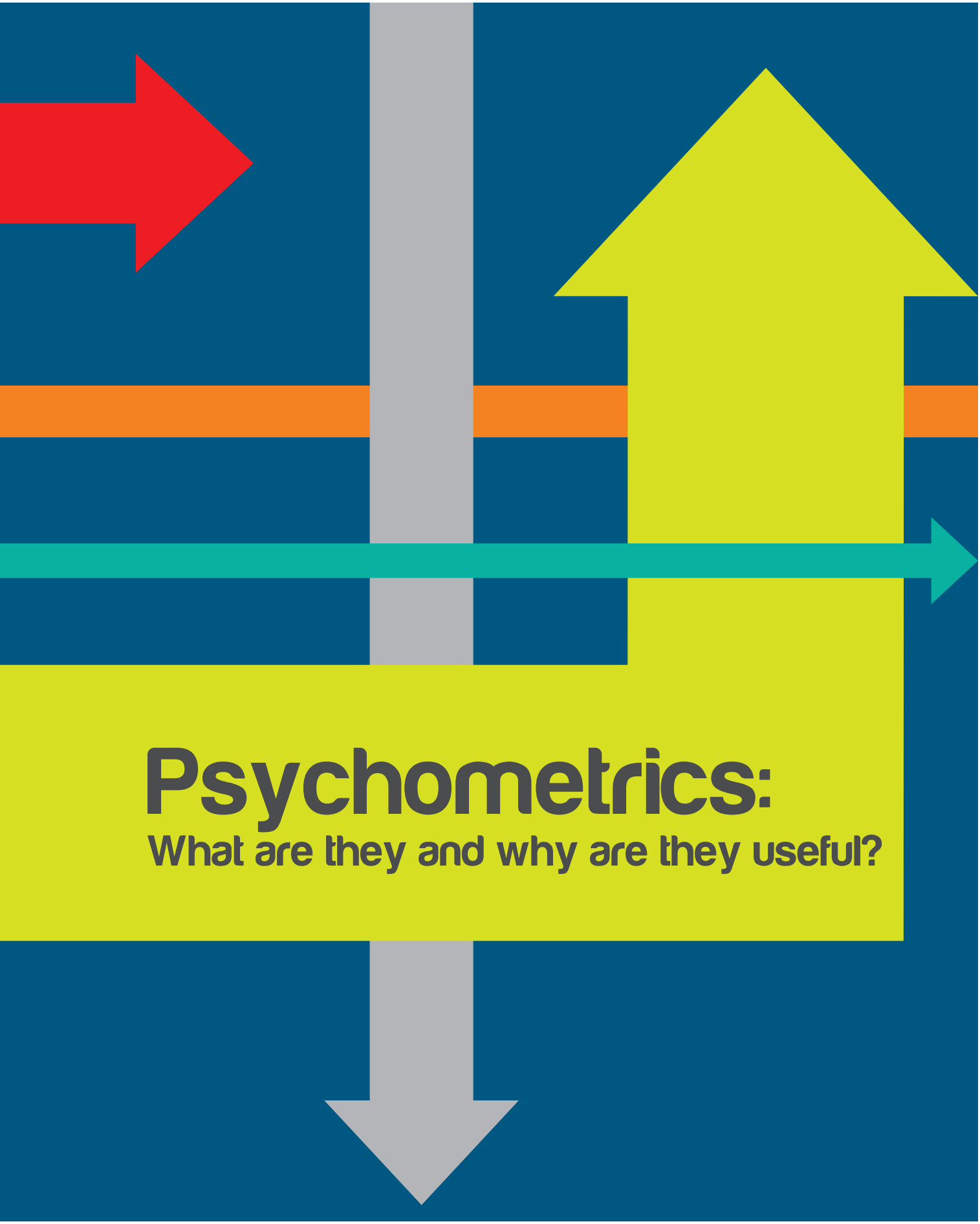
** This summary encompasses the scales in the* Online Toolbox *that map onto the skill areas that are the focus of this guide or that assess youth attitudes, behaviors or skills in related areas. Scales that typically would be viewed as indices of more distal youth outcomes are not included (i.e., scales assessing aggressive behavior on the teacher and OST program staff survey, academic competence on the teacher survey, and misconduct on the youth survey).*

Reliability and Validity of *Youth Outcome Measures Online Toolbox* Scales Reviewed in this Guide				
Scale	**Number of Items**	**Evidence of Reliability**	**Evidence of Validity**	**Corresponding Skill Area in this Guide**
Social Skills	7	Substantial	Moderate	Relationships & Collaboration
Prosocial Behavior	8	Substantial	Moderate-to-Substantial	Relationships & Collaboration
Work Habits (Teacher and Staff surveys)	6	Substantial	Limited-to-Moderate	Initiative & Self-Direction
Work Habits (Youth survey)	6	Moderate-to-Substantial	Limited-to-Moderate	Initiative & Self-Direction
Task Persistence	6	Substantial	Limited-to-Moderate	Initiative & Self-Direction
Social Competencies	7	Moderate-to-Substantial	Moderate	Relationships & Collaboration

For More Information

Kim M. Pierce
Department of Education
2038 Education (Berkeley Place)
University of California, Irvine
Irvinekmpierce@uci.edu





Psychometrics:
What are they and why are they useful?

The organization Janice works for is interested in assessing the social and emotional skills of youth who are served by the organization's after-school program and is looking for an instrument that measures these skills. After reviewing several options, she settles on an instrument that seems easy to use with questions that seem relevant for assessing the desired program impacts on youth.

Unfortunately, she encounters problems once she starts using the instrument. First, program staff seem to interpret questions very differently as they each rate a youth's skills, and there are often wide discrepancies in their ratings of a particular youth. Second, there seems to be only limited correspondence between the youths' scores on the instrument and other available indicators of their social and emotional skills, such as whether they have assumed leadership roles in program activities. These issues make Janice question whether the instrument measures youths' social and emotional skills as well as it should.[4]

The instrument Janice chose looked useful on the surface, but when it was used in the field, it was not clear that it was appropriate for the task at hand. Psychometric information might have helped Janice understand the strengths and weaknesses of the instrument before she used it.

Psychometrics are statistics that help researchers evaluate an instrument and determine if it is useful for measuring the desired concept.[5] Psychometric information can be divided into two broad categories: reliability and validity. Several different kinds of statistical evidence are used in each category to help establish that an instrument is sound.

Reliability: *The extent to which the instrument generates consistent scores each time it is used.*

One useful analogy for understanding reliability is a game of darts. If a player's darts consistently land on the same location on the board, we would say that the dart player has excellent reliability (whether or not that place is the center of the board). The same is true for research instruments that yield consistent information. Various types of reliability are discussed below.

Internal Consistency: *The extent to which the items on a scale measure the same concept.*

An *item* is a specific question or rating, and a *scale* is a set of items within an instrument that jointly measure a particular concept. For example, an instrument might include five items that are supposed to measure a youth's communication skills, and users would average or add the ratings across the five items to get an overall communication skill score. Because items forming a scale are intended to jointly measure the same concept, we can expect that the scores for each item will be related to all of the other items.[6] For example, say that the "communication" items include: (1) *How often does the youth listen appropriately to others when they are speaking?* (2) *How often does the youth express his or her ideas appropriately to others?* (3) *How often does the youth seem to have difficulty understanding what others are saying?* If the scale has high internal consistency, the rating for any one question would be related highly to the ratings for the other questions. (So if the first question received a high rating, we would expect that the second would also receive a high rating and the third would receive a low rating.) In a scale with low internal consistency, the items' ratings are unrelated to each other. Low internal consistency suggests e.g., items may not be related to each other in a meaningful way (i.e., not getting at a single underlying concept), and therefore that the overall score (the communication ability based on the average of the ratings) might not be meaningful, either.[7]

The analogy of the dartboard is useful when understanding internal consistency. Think about the individual items as the darts: The aim of the thrower is meaningless if the darts land haphazardly across the board. In the same way, an overall score like average communication is meaningless if the different items' ratings do not relate to each other. The statistic that determines internal consistency is called "Cronbach's alpha." For a scale to have acceptable internal consistency, it should be near or above the conventional cutoff of 0.70.

[4] This section on psychometrics draws heavily on a chapter written by Sean Fischer and Marybeth Shinn that appeared in the Forum's 2007 report *Measuring Youth Program Quality.*

[5] Researchers also commonly use the term "construct" to refer to the concept that is targeted by a measure. Constructs also may be referred to as the objects of measurement for scales. The constructs that are of primary interest in this guide are skills and other related attributes of youth (e.g., attitudes).

[6] In order for internal consistency to be applicable as an appropriate measure of a scale's reliability, the scale should be what researchers have called a "reflective" measure. A reflective measure is one in which it is reasonable to regard the responses to items as all emerging from (and thus "reflecting") the true level of the desired concept or construct for that youth (such as, in our example, a youth's communication skills or abilities). For this type of scale, it is expected that the responses to the different items on a scale will be consistent (i.e., very similar) because they all are (hopefully) for the most part influenced by the same thing (i.e., the underlying concept or construct). In contrast, internal consistency would not be applicable to a set of items that researchers would call a "formative" measure. A formative measure is one in which the responses to different items are each expected to help produce (or "form") an accurate assessment of the desired concept. For this type of scale, it is not expected that the responses to the different items on a scale will be consistent because each may be contributing unique and thus potentially unrelated information to measurement of the desired concept or construct. To illustrate, the sample items for the communication skills scale that we provide would be considered a reflective measure because we expect the different ratings (e.g., being skilled at listening and self-expression) to all reflect closely connected parts of the youth's underlying abilities in this area. In contrast, if the communication scale had items such as "makes speeches in classrooms" and "helps announce school-wide bulletins," we would be consider the scale to be a formative measure because we would not necessarily expect a youth who is involved in one type of specific activity (e.g., making speeches in class) to be involved in others (e.g., making school announcements). The distinction between whether a scale is best categorized as reflective or formative is not always clear cut. The large majority of the scales reviewed in this guide appear to be primarily intended as reflective measures. In the few cases where a scale appeared to be formative in its orientation, and thus internal consistency reliability would not be expected, we limited our consideration of reliability evidence to test-retest and interrater reliability. An excellent discussion of this issue can be found in an article by Bollen and Lennox (1991)[x].

[7] Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110, 305-314.

Interrater Reliability: *The extent to which raters agree when evaluating the same youth at the same time.*

When an instrument involves observers providing ratings, it is also important to consider interrater reliability. For accurate assessment, an instrument should yield consistent scores regardless of the idiosyncrasies or tendencies of individual raters. When findings depend largely on who is doing the rating (e.g., if Rater A is more likely to give favorable scores than Rater B), it will be difficult to get a reliable sense of a youth's true level of skill or ability. For this reason, organizations should consider the interrater reliability of an instrument even if only one rater will be rating each youth. Poor interrater reliability often stems from ambiguous questions that leave a lot of room for individual interpretation, and such ambiguity is not always immediately apparent from looking at the items on the instrument.

Some instruments' developers offer training for raters. If you cannot receive formal training on an instrument, it is still desirable whenever feasible to train raters yourself before conducting an assessment or evaluation. Organizations can hold meetings to review each question individually and discuss what criteria are necessary to assign a score of 1, 2 or 3, etc. If possible, raters should go through "test cases" to practice using the instrument. When disagreement occurs on individual questions, raters should discuss why they chose to rate a youth the way they did and come to a consensus. Practice evaluations will help raters develop a mutual understanding of what to look for so that they can rate youth in a similar manner.

Several statistics are available to measure interrater reliability. A simple percentage agreement is perhaps the most straightforward of these statistics. It does not account for those instances in which raters agree simply by chance, however, and for this reason is less preferred than alternative statistics such as kappa and the intraclass correlation. These methods also allow for more than two raters to be considered in the interrater reliability statistic. For this guide, we considered findings to be relevant to interrater reliability when the raters are observing the youth in the same setting at generally the same point in time. This generally involved either two different OST program staff or two of the youth's teachers providing ratings of the youth. Otherwise, it was assumed that factors other than the measure's inherent lack of reliability could be resulting in differences in scores across raters, such as a youth exhibiting a different level of social skills when in an OST program than when at school.

Test-retest Reliability: *The stability of a scale's scores over time.*

If a youth's scores on a scale differ very little across two different times of measurement, it has strong test-retest reliability. In general, test-retest reliability is a meaningful form of reliability only when the measurements occur over a short enough period of time for the youth's skills to have not changed due to reasons such as normal development or participation in a program.

Let's return to our earlier example of a scale that measures communication skills. If the scale was completed twice by a group of youth over an interval only a few weeks, it would be reasonable to expect the same youth to receive relatively the same scores each time. In this report, we consider findings to be relevant to test-retest reliability only when the interval between measurements is three months or less. Typically, test-retest reliability is assessed using either the Pearson correlation coefficient or an intraclass correlation. For the measures reviewed in this guide, the Pearson correlation coefficient was used in all instances to assess test-retest reliability. For this statistic, a value of .70 or greater often would be considered to indicate an acceptable level of reliability.[8]

[8] In some cases, the average score for a group of youth on a measure may tend to increase or decrease across two administrations of the measure. In this case, if the relative standing (rank-ordering) of youth on the measure is relatively unchanged, the Pearson correlation coefficient will still tend to indicate a high level of test-retest reliability. In comparison, the intraclass correlation can be useful if there is an interest in also detecting whether youth tend to receive the exact same score on a measure across administrations.

Validity: ***An instrument's ability to measure what it is intended to measure.***

If a scale on an instrument is supposed to measure a youth's skills in a particular area, then it would be valid if it yielded accurate information about the youth's abilities in that area. The game of darts again provides a useful analogy. Whereas reliability is about the player consistently throwing darts to the same location, validity relates to whether or not the player is hitting the bull's eye. The bull's eye is the concept or construct an instrument is intended to measure. Although reliability is essential, it is also important to know if an instrument is valid. (Dart players who consistently miss the board entirely may be reliable – they may hit the same spot over and over – but they are sure to lose the game.)

Sometimes an instrument may look like it measures one concept when in fact it measures something different or measures nothing particularly well. For example, returning again to our example of a scale that claims to measure communication skills, such a scale would not be particularly valid if it focused solely on whether youth liked to talk a lot.

Validity can be challenging to assess because the concepts of interest are often not tangible or concrete. Unlike the case of reliability, there is no specific number that tells us about validity. Rather, validity is more of a qualitative assessment that is arrived at by considering the preponderance of available evidence. Several different types of statistical analyses that can be used to inform judgments about a measure's validity are discussed below. These analyses have been associated with different types of validity, the names for which are also provided below. It is important to remember, however, that ultimately all of the analyses share the same goal of helping us to judge how well the scores on a scale capture whatever it is intended to measure.

It also is important to keep in mind that assessments of a scale's validity should always be linked to the particular intended use of the measure. Consider, for example, two scales that each have published evidence of being valid measures of problem-solving ability. In deciding which measure to use in the evaluation of an OST program, it would be appropriate to consider which scale is likely to provide the most valid assessment of the particular aspects of problem-solving ability that the program is intended to improve. If the program has the goal of strengthening problem-solving skills for resolving conflicts with peers, for example, then the scale that appears most likely to be valid for assessing these aspects of problem-solving ability would be the most appropriate choice. Ultimately, then, judgments of a scale's validity can not be made in a vacuum, but rather must be informed by careful consideration of the specific purpose or goal for which a measure will be used.

Convergent Validity: ***The extent to which the scores on a scale are associated positively with scores on scales that measure the same or highly similar concepts.***

If two scales are presumed to measure the same or similar concepts, one would expect scores on the two scales to exhibit a high level of agreement or overlap. For example, suppose researchers have developed a new scale (Scale A) that is intended to measure youths' teamwork skills. To assess its validity, researchers might administer both Scale A and another scale (Scale B), which is already well-established as a valid measure of teamwork skills, to the same youth. Assuming that Scale A is also a valid measure, we can expect that when Scale B finds that a youth has good teamwork skills, Scale A will as well. If this is not the case, we would conclude that Scale A probably does not adequately measure teamwork skills.

Unfortunately, in practice, assessments of convergent validity can be complicated by several considerations. One common challenge is finding a scale that has well-established validity as a measure of whatever concept the scale of interest is supposed to measure. As we have already noted, assessments of validity are not cut and dried. Even in the most ideal of circumstances, we are unlikely to ever be able to conclude that a scale's

[9] It is useful to keep in mind, too, that claims of validity are population-specific. In other words, just because a measure such as Scale B has been indicated to be valid for assessing a concept like teamwork in one population (e.g., youth ages 14 and older), this does not guarantee it will be valid for all other populations (youth ages 10-12). Accordingly, if researchers are seeking to validate Scale A for use with a population of youth (or raters) that is different from the one(s) on which Scale B has been validated, a lack of expected association between scores on the two scales could reflect limitations in Scale B's validity for the new population more than it does a lack of validity for Scale A.

validity is established with absolute confidence. With this in mind, returning to our example above, suppose Scale A does not show a strong association with Scale B. Is this because Scale A is not a valid measure of teamwork, or is it a reflection of limitations in Scale B's validity?[9] Another important consideration is the well-established tendency of data on a set of scales that are collected from the same informant or using the same method (such as youth self-report or teacher ratings) to show overlap for reasons other than the different scales involved assessing the same concept. For example, an observer might tend to rate the same youth relatively high or low across two areas, even if the youth's abilities or skills differ across those areas, because of what has been called a "halo effect." For this reason, researchers typically give more weight to convergent validity evidence that comes from different informants or methods (e.g., if Scale A is a self-report measure of teamwork and Scale B is based on ratings by the staff of an after-school program).

Discriminant Validity: ***The extent to which scores on scales that measure distinct concepts are not associated at unexpectedly high levels.***

If two scales are presumed to measure different concepts, one would not expect scores on the two scales to exhibit a strong association. Let's continue with the same example of a new scale (Scale A) that is supposed to measure teamwork. Researchers might administer this scale to a group of youth along with another scale (Scale C), which is a well-established measure of a concept that is distinct from teamwork, such as creativity. If Scale A is a valid measure, we can expect that the scores from Scale C will not exhibit a strong relationship with scores from Scale A. If this type of strong relationship were found, we would have reason to question whether Scale A is a valid measure of teamwork skills.

But just how strong of an association between scores on the two scales would be so high that it could cast doubt on the Scale A's discriminant validity? To help address this question, it is useful to have some type of benchmark available. One benchmark used by researchers would be the level of association that Scale A shows with another established measure of the same concept. This would include a scale such as Scale B, the scale that we referred to above in discussing assessment of a scale's convergent validity. In general, if Scale A has discriminant validity, we would expect that its association with Scale C would be less strong than its association with Scale B.

The same factors that we noted can complicate assessments of convergent validity can also make it challenging to gauge a scale's discriminant validity. Suppose, in our example above, that Scales A and C are both based on the self-reports of youth, whereas Scale B is based on ratings of teachers. Scores on Scale A could be associated with those for Scale C simply because both scales come from the same source (something researchers refer to as "shared method variance"). This association could be stronger than Scale A's association with Scale B, thus suggesting that the scale's discriminant validity is low even though this may not be the case. To help sort out these kinds of issues, it is best to have available what researchers call "multitrait-multimethod" data, in which multiple concepts are each measured using multiple methods. In our example, this could involve adding a fourth measure, Scale D, that assesses the same concept as Scale C (creativity) but does so based on teacher ratings. Among other things, this would allow us to see if discriminant validity of Scale A is supported by it having an association with Scale B (teamwork assessed using teacher ratings) that is less strong than its association with Scale D (creativity assessed using teacher ratings). This type of comparison is desirable because neither association will be influenced or biased by shared method variance.

Criterion Validity: ***The degree to which a measure is related in expected ways to some type of criterion or outcome, measured either at the same time (concurrent validity) or a later time (predictive validity).***

If a scale does a good job of capturing the concept that it is intended to measure, then scores on the scale would be expected to be related to criteria or outcomes that are influenced by that concept. For example, if a scale is supposed to measure the abilities of youth to persist on difficult tasks, then we would expect that youth who receive higher scores on the measure would also be more successful in school.

There are two types of criterion validity: concurrent validity and predictive validity. With concurrent validity, the scale and the criterion or outcome are measured at the same time. With predictive validity, the scale is measured at one point in time, then the criterion or outcome is assessed at a later point in time. Thus, if youth who score higher on the scale intended to measure task persistence are found to also be earning higher grades in school at the same point in time, this would be support for concurrent validity. If these youth also were found to be more likely to graduate from high school at some point in the future, this would indicate predictive validity. Typically, greater weight and significance are attached to predictive validity evidence. This type of evidence is especially well-suited to assessing whether scores on a scale demonstrate expected associations with outcomes that may emerge only at later points in a youth's development, such as educational attainment or involvement in certain types of problem behavior.

Researchers may use both theory and prior research findings to determine which outcomes are most appropriate to establish criterion validity. Ultimately, these determinations are judgment calls subject to debate and disagreement. A further complicating consideration is the potential for the outcome or criterion measure to have limited validity, which could then be an alternative explanation for why the scale of interest does not predict that measure.

Construct Validity: ***The degree to which a measure is related in expected ways to measures of hypothesized antecedent and consequent concepts, ideally within a defined model.***[10]

Typically, the concept that is supposed to be measured by a scale can be expected to not only have an effect on other concepts, as just discussed with criterion validity, but also to be influenced by different concepts as well. There are typically many potential influences on whatever is intended to be measured by a scale. One important type of influence for the measures reviewed in this guide would be participation in an OST program. Many OST programs, for example, are intended to provide youth with positive learning and mastery experiences. It is reasonable to expect that participation in such programs should, among other possible outcomes, strengthen the abilities of youth to show sustained effort when faced with difficult or challenging tasks. Accordingly, program participation should lead to higher scores on a measure of task persistence like the one we referred to above.

OST program participation, of course, is only one of many factors that could be predicted to influence scores on a measure intended to assess abilities in this area. We might also expect, for example, that youth who experience difficulties with attention or hyperactivity would find it more difficult to persist on tasks and thus score lower on the scale. Here, too, theory and prior research findings help researchers determine which antecedent concepts are most appropriate to examine for a given scale. Ideally, there will be a well-delineated model available that depicts an integrative network of relationships between several different antecedent concepts, the concept of interest and potential consequents or outcomes (i.e., concepts expected to be influenced by the concept of interest). Specialized methods, most notably structural equation modeling, are available to test whether data collected on a set of relevant measures provide support for a proposed model or theory. For purposes of informing assessment of a scale's construct validity, we would be most interested in the parts of the model that involve the scale's linkages with measures of concepts that are expected to either influence or be influenced by the concept the scale is intended to measure.

If findings are consistent with theoretical predictions for a scale, we would conclude there is support for a scale's construct validity. If findings are not consistent with what is expected, this could indicate an issue with a scale's validity. Alternatively, the same results could just as easily indicate a problem with the accuracy of the associated theoretical predictions. Consider, for example, a situation in which participating in an OST program is not found to lead to higher scores on our hypothetical scale intended to measure task persistence, even though theory suggests that the program should improve skills in this area. Determining whether the reason for this finding is a lack of validity for the scale (the program does improve task

[10] Researchers sometimes use the term construct validity more broadly to encompass all different types of validity evidence that are available for a measure. [11] To complicate matters further, it also could be the case that the OST program was poorly implemented. This could be another reason for the unexpected results in our example rather than a problem with either the scale's validity or the theoretical prediction about what outcomes are affected the program when it is implemented appropriately.

persistence, but the scale is not able to detect its effects on this outcome), a problem with our theoretical prediction (the program as designed does not have an effect on task persistence), or perhaps both of these reasons is not a simple undertaking.[11]

Generally speaking, in this type of situation it is advisable to look to additional sources of information for guidance. This could include whether the scale has exhibited good convergent validity with other well-validated measures of the same concept, in which case we would tend to question the accuracy of our theoretical model more than the validity of the scale. We also could look at whether the same theory has received support when using other scales to assess the concept of interest, in which case we then would be more likely to question the validity of the scale.

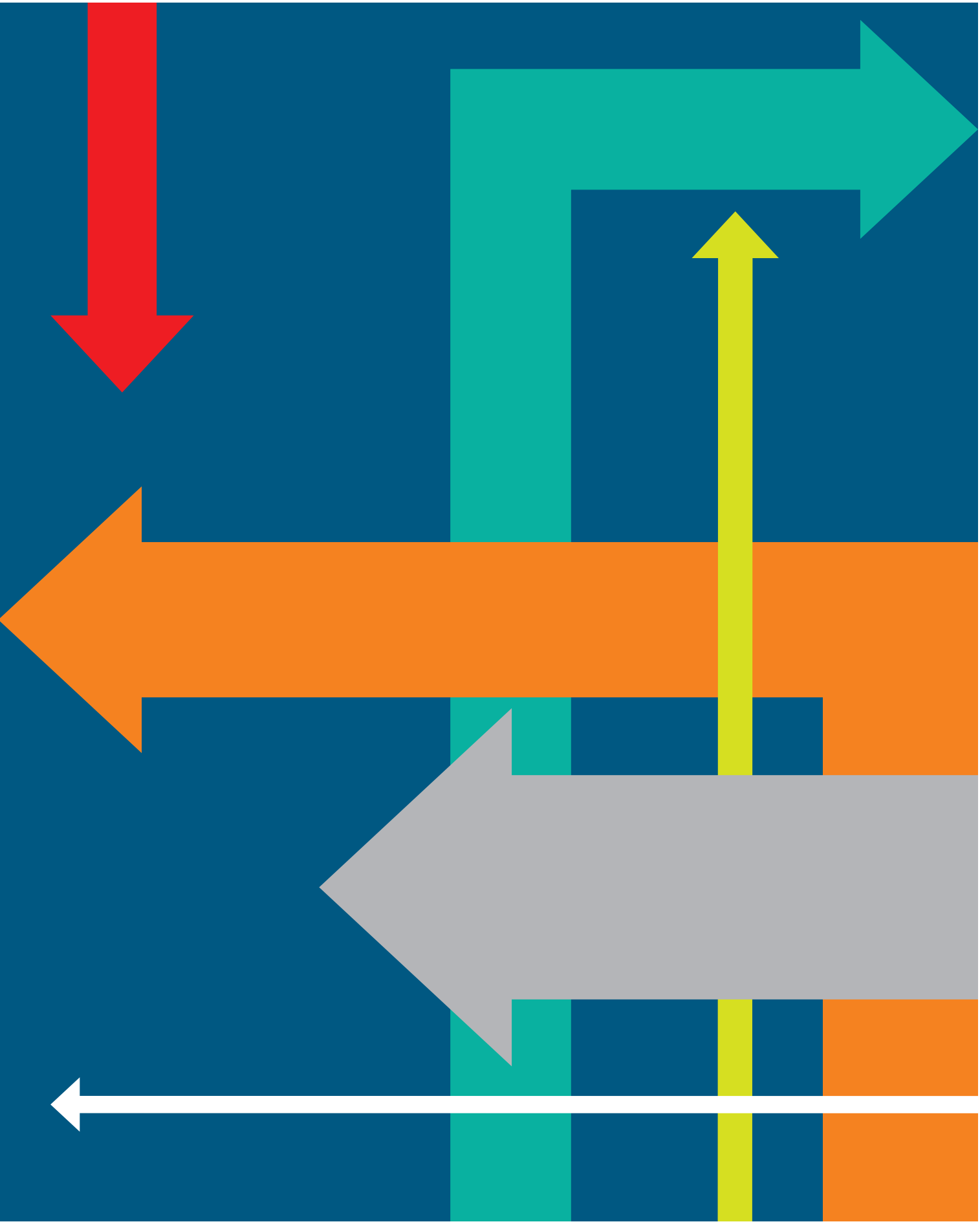
Validity of Scale Structure: ***The extent to which individual items on an instrument measure the different concepts that the instrument is intended to assess. (This is appropriate only for instruments that have divided their items into scales.)***[12]

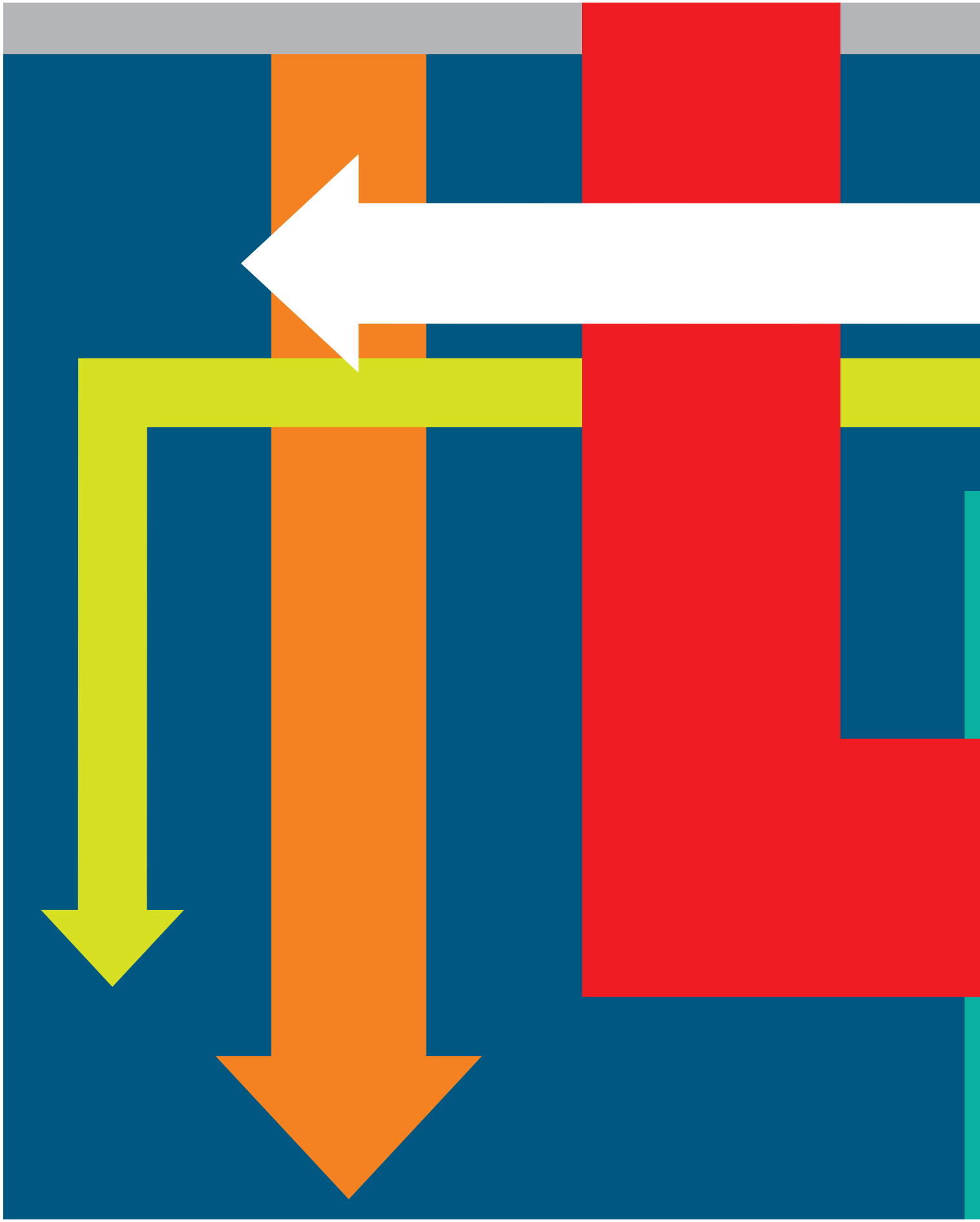
As already stated, scales are composed of several items that, when averaged or summed, create an overall score for a specific concept. Often, the items on a single instrument will be used to derive several different scales, each intended to measure a different concept. The validity of scale structure is important is because we want to know whether the items on an instrument have been grouped appropriately for purposes of computing scales that represent the different concepts that an instrument seeks to measure. Determining whether the individual items on an instrument adequately measure the concepts they are intended to measure can be difficult, although conducting what is called a factor analysis is one helpful way to do so. Factor analysis examines which items are similar to each other and which are different, and helps address whether certain groups of items can be assigned to the same scales within an instrument. Ideally, these groupings will correspond to the instrument developer's hypotheses or assumptions.

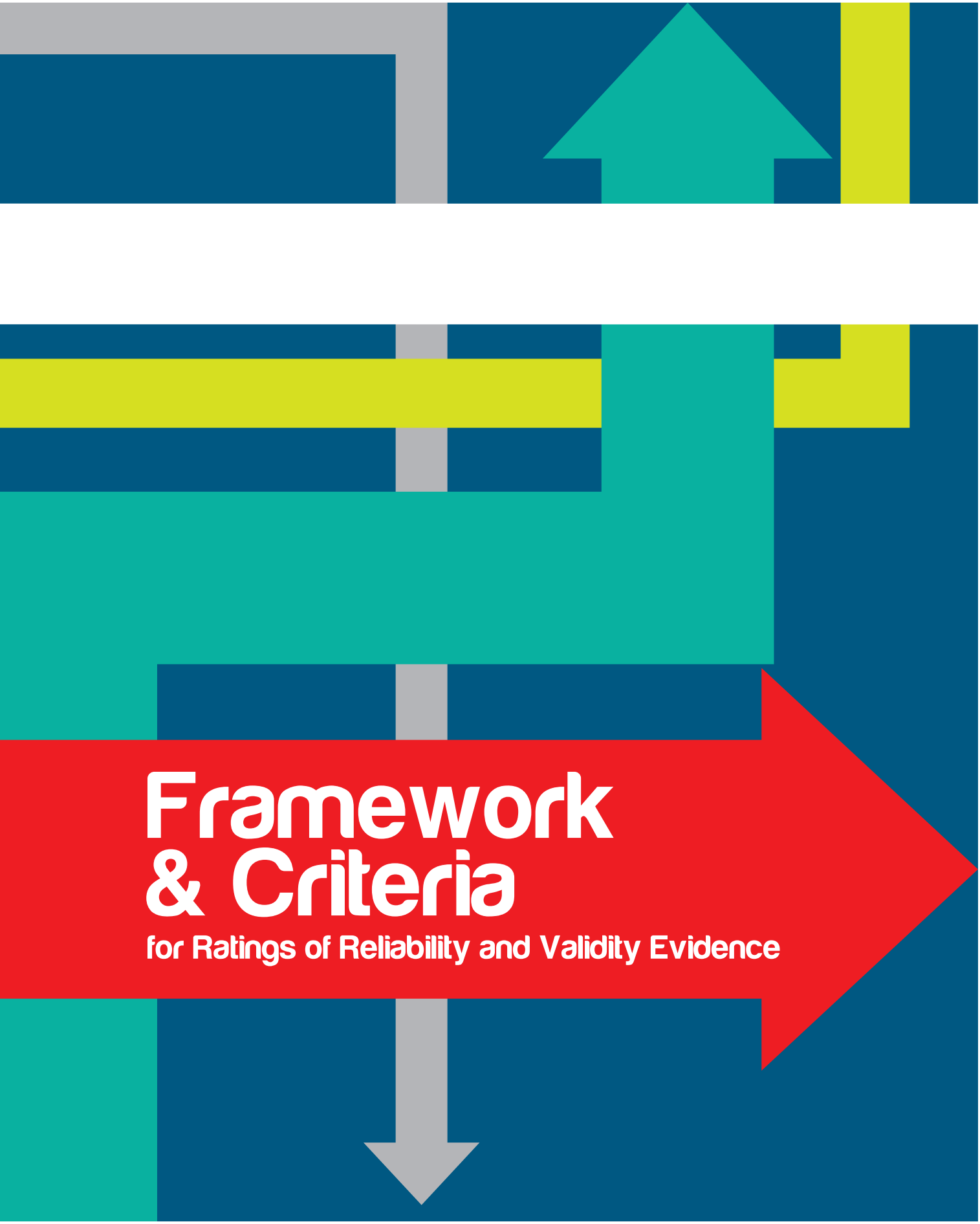
For example, imagine an instrument with two scales intended to assess skills in the areas of Task Persistence and Task Management. Suppose that in nearly all cases where youth receive high ratings on the items that make up the task persistence scale, they also receive similarly positive ratings on the items that make up the task management scale. Because of the high degree similarity in ratings for the two sets of items, a factor analysis likely would indicate that the items involved are actually measuring just one concept, not two. In this case, it could make more sense to compute just one scale from the items involved, perhaps renamed Task Completion.

Factor analysis can also help determine if a scale on an instrument actually incorporates more than one related concept. Imagine that we have an instrument with a scale called Social Academic Problem Solving, but that a factor analysis finds responses to the items on the scale are not all closely related. This would suggest that the items assigned to the scale are not all measuring the same concept. We might discover, for example, that some items relate to Social Academic Problem Solving, whereas another set relates to Problem Solving with Teachers. Ideally, when findings of a factor analysis suggest revisions to how an instrument is scored, the results are confirmed through analyses conducted with a new set of data. The technique of testing support for a particular hypothesized scale structure for the items on an instrument is called confirmatory factory analysis.

[12] In this guide, validity of scale structure is considered primarily when evaluating the psychometric properties of the overall instruments that include the scales that are reviewed. Where appropriate, however, in reviewing individual scales we consider whether factor analysis findings support the distinctiveness of the items that are used to compute the scale relative to those assigned to other scales (and thus intended to assess other concepts) on the same instrument because of the relevance of this evidence for assessing the scale's discriminant validity.







Framework and Criteria for Ratings of Reliability and Validity Evidence

This is an overview of the procedural steps and guidelines that were used to arrive at the ratings of reliability and validity evidence reported for each of the scales reviewed in this guide.[12] An overview of the framework used is shown below (Figure 2).[13] Those interested can request a copy of the complete rating system from the authors. There are inherent limitations to any effort 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 and to always consider the ratings that are provided for a scale in conjunction with the narrative accounts of the underlying evidence.

Figure 2: Overview of Framework for Ratings of Reliability and Validity Evidence

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

***a* 1**=*None;* **2**=*Limited;* **3**=*Moderate;* **4**=*Substantial;* **5**=*Extensive.*
***b* 1**=*Poor;* **2**=*Fair;* **3**=*Good;* **4**=*Very Good;* **5**=*Excellent.*
***c* 1**=*None;* **2**=*Limited;* **3**=*Some;* **4**=*Most;* **5**=*All or Nearly All.*
***d* 1**=*Very Low;* **2**=*Low;* **3**=*Moderate;* **4**=*High;* **5**=*Very High.*
***e* 1**=*Highly Inconsistent;* **2**=*Moderately Inconsistent;* **3**=*[No Label];* **4**=*Moderately Consistent;* **5**=*Highly Consistent.*

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?

Having answered these questions, we next evaluated the available evidence as it pertained to each of several different facets of reliability and validity (see Figure 1). The the section on psychometrics (see page 51) of this guide includes a brief explanation of each of these types of reliability and validity. Orienting questions similar to those listed above were used to facilitate ratings of the available evidence as it related to each facet of a scale's reliability and validity. In the case of reliability, these questions were used to identify which facets 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 2). Guidelines were developed to facilitate the assignment of these ratings for different facets of reliability and validity. For example, for rating 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. In assessing the quality and rigor of evidence for criterion-related validity, for example 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

[13] 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 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. According, the assessments that are provided should be regarded as having the potential to be broadly informative only.

[14] 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 ***C****ompendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions* prepared by Mathematica Policy Research, Inc. (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).

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 aspect 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 aspects 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 not supportive.

The final step in the process was 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, corresponds to a rating of 3, and an assessment of "Moderate-to-Substantial"corresponds 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. For example, 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 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 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 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 inter-rater reliability was not conducted. Furthermore, the validity of the rating system itself has not been evaluated. 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.

[i] Metlife (2011). *The Metlife Survey of the American Teacher: Preparing Students for College and Careers.* Metlife: New York, NY. www.metlife.com/about/corporate-profile/citizenship/metlife-foundation/metlife-survey-of-the-american-teacher.html?WT.mc_id=vu1101

[ii] The Conference Board, Inc., the Partnership for 21st Century Skills, Corporate Voices for Working Families, and the Society for Human Resource Management. (2006). *Are They Really Ready for Work?* http://www.p21.org/documents/FINAL_REPORT_PDF09-29-06.pdf

[iii] Murnane, R. and Levy, F. (1997). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy.* New York, NY: Free Press.

[iv] Durlak, J. A., & Weissberg, R. P. (2007). *The impact of after-school programs that promote personal and social skills.* Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

[v] Silva, E. (2008). *Measuring Skills for the 21st Century.* Education Sector: Washington, DC. Should be associated with same page, same paragr, last sentence ending in "…still evolving."

[vi] The Secretary's Commission on Achieving Necessary Skills, U.S. Department of Labor (June 1991). *What Work Requires of Schools: A SCANS Report for America 2000.* U.S. Department of Labor: Washington, DC.

[vii] Yohalem, N. and Wilson-Ahlstrom, A. with Fischer, S. and Shinn, M. (2009, January). *Measuring Youth Program Quality: A Guide to Assessment Tools, Second Edition.* Washington, DC: The Forum for Youth Investment. http://forumfyi.org/content/measuring-youth-program-quality-guide-assessment-tools-2nd-edition.

[viii] Denham, S., Ji, P., and Hamre, B. (October 2010). *Compendium of Preschool Through Elementary School Social-Emotional Learning and Associated Assessment Measures.* Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

[ix] Vandell, D. L., O'Cadiz, P., Hall, V., & Westover, T. (2010). *California After School Outcome Measures Project: Phase II Final Report.* Report to the David and Lucile Packard Foundation.