MEASURING SOFT SKILLS & LIFE SKILLS IN INTERNATIONAL YOUTH DEVELOPMENT PROGRAMS

A REVIEW AND INVENTORY OF TOOLS

MAY 2017

This report is made possible by the support of the American People through the United States Agency for International Development (USAID) under YouthPower Action, Contract number AID-OAA-TO-15-00003/AID-OAA-I-15-00009. The contents of this report are the sole responsibility of FHI 360 do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

This page is intentionally left blank

YOUTHPOWER ACTION

Measuring Soft Skills & Life Skills in International Youth Development Programs: A Review and Inventory of Tools

May 2017

Disclaimer:

The author's views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

This report is made possible by the support of the American People through the United States Agency for International Development (USAID) under task order contract number AID- OAA-TO-15-00003, YouthPower Action under IDIQ contract number AID-OAA-I-15-00009, YouthPower: Implementation.

Recommended format for citation: Galloway, T, Lippman, L., Burke, H., Diener, O., and Gates, S. (2017). Measuring Soft Skills & Life Skills in International Youth Development Programs: A Review and Inventory of Tools. Washington, DC: USAID's YouthPower Implementation IDIQ- Task Order 1, YouthPower Action.

Photo Credit (bottom right, cover page): FHI 360/USAID El Salvador, Education for Children and Youth Program

TABLE OF CONTENTS

FIGURES

TABLES

ACRONYMS AND ABBREVIATIONS

BARS	Behaviorally anchored rating scale
CAWS	Child and Adolescent Wellness Scale
CHKS	California Healthy Kids Survey
CPYDS	Chinese Positive Youth Development Scale
PISA	Program for International Student Assessment
RSQ	Responses to Stress Questionnaire
SEHM	Social and Emotional Health Module
SENNA	Social and Emotional or Non-Cognitive National Assessment
SJT	Situational judgment test
SRH	Sexual and reproductive health
USAID	United States Agency for International Development
WfD	Workforce development

1. Executive Summary

PURPOSE

In recent years, as the evidence base on the importance of soft and life skills for fostering positive youth outcomes has grown, international youth development programs have increasingly focused on interventions that develop those skills (also referred to as socio-emotional skills, transferrable skills, non-cognitive skills, and developmental assets, among other terms). For the sake of clarity, this paper hereafter uses the term "soft skills" to refer to this body of skills -- as this term is widely understood among youth, employers, program implementers, and researchers – while acknowledging that the term "life skills" is preferred by many in the sexual and reproductive health field, and that other terms may be preferred in other contexts. The growth in soft skills-focused interventions has resulted in an urgent need among youth development programs for soft skill measures that can be used for program implementation and evaluation.

Soft skills measurement is still an emerging area of research, however, and the landscape of soft skills measures is varied and fragmented across disciplines. This report attempts to bring clarity to this field by identifying existing instruments that can be used or adapted for use across youth programs in developing country contexts. USAID's YouthPower Action project has completed a review of soft skill measurement tools and created an inventory describing characteristics that can be useful to international youth development programs that seek to assess participants' soft skills. This summary report describes general findings about the universe of tools reviewed, as well as specific findings about tools that measure a select set of key soft skills, and suggests recommendations for improving those resources.

Prior work by YouthPower Action identified key soft skills that foster positive workforce, violence prevention, and sexual and reproductive health (SRH) outcomes (see the papers "Key Soft Skills for Cross-Sectoral Youth Development" by Gates et al. (2016) and "Key 'Soft Skills' that Foster Youth Workforce Success: Toward a Consensus Across Fields" by Lippman et al. (2015)). From those systematic reviews of the literature, three skills emerged with the highest degree of research support across all three outcome areas: self-control, positive self-concept, and higher order thinking skills. Four additional skills rose to the top for certain outcome areas, but not all: social skills, communication, goal orientation, and empathy. These cross-cutting seven skills were the focus of the measurement tool review. In addition, the skills of hard work and dependability, responsibility, and positive attitude were also noted in the search for measures since they received strong support in the workforce literature, and the latter two received support at a lower level across all three outcome areas.

State of the Field of Soft Skills Measures

Measurement tools may be used for a number of different purposes for international youth development programs, including: 1) formative assessments, to inform program participants of their progress; 2) implementation, to provide programs with information for the purpose of better

implementing their programs; 3) summative or descriptive, to describe or monitor the progress of youth at the group-level within a program; and 4) evaluative, to evaluate the effectiveness of a program in developing skills or having an impact on specific outcomes through skills development. Measures may differ by their use in form, content, the nature in which scores are reported, and the level of standards applied (Stecher and Hamilton, 2014). Multiple types of measurement tools exist, and there are a number of different ways to organize them. The inventory of measures includes: self-reports and self-ratings, and ratings and observations by others; performance assessments, direct assessments from tests, and simulations, including games; and mixed methods measures.

The field of soft skills measures faces a number of methodological challenges, including:

- Balancing technical considerations such as reliability, validity, and measurement invariance
- Using tools to reliably measure change in skills over time, when measuring a soft skill at a single point in time is itself challenging
- The prevalence of self-report methods that are known to suffer from biases
- Developing or adapting tools for use across cultures and contexts with limited resources
- Lack of implementer inclusion in tool design

Balancing these challenges can be difficult and there are trade-offs for every method. This report reviews these challenges and discusses potential solutions.

Methodology

The YouthPower Action team conducted a review of close to 300 instruments to inform the field. Instruments were screened out that did not address the key soft skills, were not developed for youth between the ages 12 and 29, or which had a cost associated with their purchase and administration. Free access to instruments was considered necessary for programs around the world to use them and to build the state of the evidence for the field. Seventy-four instruments met those three criteria. An inventory of those measures was then created, which addressed characteristics of each instrument, which are described in detail in the Methodology section of the paper.

The team then reviewed each tool based upon a set of criteria that was developed with input from soft skill measurement experts and implementers. Each of these criteria are described in more detail in the Methodology section. Each tool was then scored tool according to the degree to which it met a set of seven criteria. The criteria include:

- Evidence of use by international youth development programs
- Evidence of validity
- Relevant validation sample
- Used with youth development outcomes of interest
- Evidence of reliability
- Evidence of international usage

- Ease of administration (points were granted for not needing trained personnel for administration, short length, and availability in other languages)

The tools were then divided into three groups based upon the degree to which they met all of the criteria: high (meeting five to seven criteria), medium (meeting from three to fewer than five criteria), and low (meeting fewer than three criteria).

Limitations of Methodological Approach

It is important to note a few limitations of the methodology. First, although this project has identified many key soft skills measurement tools, it should not be considered a comprehensive list. Second, this identification and screening of tools represents those tools that were available as of 2016. Tools may have been excluded due to their incomplete nature, or because they are still undergoing validity and reliability testing, or they fell outside the scope of work for this project.

Third, comparisons of tools are difficult, which underlines the importance of using or adapting tools for specific purposes and contexts. Although the focus of this project is on tools that would be appropriate to contexts in which USAID and other international youth development efforts are working, many tools identified target U.S. or Western educational contexts, reflecting the burgeoning interest in soft skill measurement.

Finally, given the breadth of contexts in which soft skills are measured, it should be apparent that no one tool is capable of meeting all the measurement needs of youth development programs. The purpose of the list of tools that have been categorized in the inventory is to describe the breadth and depth that current tools reach both in the soft skills they measure and in the potential uses they may serve. The findings should not be used to definitively mark one tool as more useful or "better" than others; instead the scores are meant to describe differences among the tools with respect to measuring key cross-cutting soft skills for youth. Their usefulness will vary according to the needs of each program, including the skills that are the focus of the programs, the age group participating, and the purpose for which the tool will be used.

Overall Findings from the Inventory

High-scoring tools exist for each of the key soft skills previously identified. The evidence compiled suggests that the field of measurement is generally well-aligned with literature on the key soft skills that are most supported by evidence as promoting positive cross-sectoral outcomes. For example, self-control has the most measures in the inventory (43), whereas positive attitude has the least (19). Overall, the field of measurement remains largely dominated by self-report measures for most of the key skills, and the availability of other types of measurement (e.g., report by others) is limited and uneven.

Although evidence of acceptable levels of reliability were found among the majority of tools (63 percent), evidence of acceptable levels of validity was found among a minority of tools (44 percent). The age group that enjoyed the most tools was 15–19 years, followed by 12–14.

There were fewer tools available in older age groups. Generally, the majority of tools met criteria for ease of administration. A number of tools have been tested in various regions of the world.

Selected High-Scoring Tools Measuring Top Three Skills

The final step was to highlight the tools that earned high scores *and* measured the top three skills that are linked to all three outcomes areas: higher order thinking skills, positive self-concept, and self-control. These tools were selected from the larger inventory as potentially of greatest interest to international youth development programs working to promote positive workforce and sexual and reproductive outcomes, and preventing violence. Some of the tools have been used in conjunction with other outcome areas as well, such as education, psychological and emotional health, substance abuse, and health (see the inventory for more details).

There are 10 such tools (see Table 5 on page 42 for a breakdown of skills measured by each tool):

- California Healthy Kids Survey: Social and Emotional Health Module
- Chinese Positive Youth Development Scale (CPYDS)
- SENNA 1.0
- SENNA 2.0
- Child and Adolescent Wellness Scale
- The Anchored BFI Tool
- The Big Five Inventory
- Knack
- Jamaica Youth Survey
- Responses to Stress Questionnaire (RSQ)

In two cases, this group includes different versions of the same tool (SENNA 1.0 and 2.0; the Big Five Inventory and the Anchored BFI).

In the this report, each of these tools is described in depth from the perspective of their utility for international youth development programs.

The tools generally fall into the following three categories of usage:

***Program evaluation*:** The Chinese Positive Youth Development Scale and the Jamaica Youth Survey meet the above-mentioned criteria and have been used to evaluate international youth development programs. The Chinese PYD Scale has the advantage of assessing eight of the top nine skills, whereas the Jamaica Youth Survey assesses five.

Group performance monitoring: The California Healthy Kids Survey, Social and Emotional Health Module, and the Brazilian SENNA surveys, are instruments of excellent quality that are useful for monitoring group performance for summative, descriptive purposes, and which have been used in schools and school districts. They could be used for evaluations where group-level data are needed, but they are not validated for use by evaluations that seek to measure individual improvements in soft skills over a program's duration.

Individual assessments: The rest of the tools can be used for individual psychological or skill assessments and have been shown to be correlated with outcomes of interest. They can be used in formative assessments in which program staff give feedback and coaching to youth participating in the programs, and when grouped, may be informative for improving the targeting of skills within a program and for program implementation purposes. They are useful for detecting differences among individuals in a program at one point in time, but they may not be sensitive enough or validated for evaluation designs that need to detect improvements in individuals' skills over the duration of a youth development program.

Programs will need to evaluate the tools in this inventory and this extracted list of tools for their own purposes. A measurement instrument needs to align with the program it is being used for, as well as the design of an evaluation. Considerations may include whether the skills being addressed by the program match the skills that are measured in the assessment under consideration, whether the tool has been validated for use with youth of the same age as are in the program, whether it enjoys acceptable levels of validity and reliability, whether the tool has been used for the same purpose as is envisioned by the program or program evaluation, and whether it has been used to measure an impact on outcomes of interest to the program. Not all criteria will be of equal importance to every program.

Challenges for the Field

Many excellent tools measure soft skills and new ones are being developed. In general, the field currently exhibits some weaknesses and limitations that obstruct their usefulness for program monitoring and evaluation. In addition, some challenges affect the ability to build evidence in the field *across* programs, which is essential in order to learn what is working and which programs need to be scaled up. Several challenges need to be addressed by the field.

Terminology: The lack of a common terminology and skill definitions across measurement instruments hampers the ability of program implementers and evaluators to choose instruments that match the set of skills addressed by programs, and to compare results across programs. It also hampers the ability to build the evidence across countries, cultures, research disciplines, policymakers, funders, and practitioners. Proposed common terminology and skill definitions that would bring coherence to the field were suggested in "Key 'Soft Skills' that Foster Youth Workforce Success: Toward a Consensus Across Fields" (Lippman et al., 2015) which was drawn from the research terminology across fields and studies, but also with attention to the terms used by youth, practitioners, and employers.

Evidence of reliability and validity: As noted in the analysis, many tools lacked evidence of reliability and validity, as well as differential item functioning and measurement invariance, which are essential to provide confidence in the tools. Developers need to be encouraged to publish the results of their tests with their validation samples, and those who have used the tool for assessing youth along with outcomes need to be encouraged to report their reliability and validity.

Prevalence of self-report methods: All of the 10 tools highlighted above—except the Knack game—and most of the tools in the inventory use youth self-rating scales, which suffer from

reference and social desirability biases. It is known that there is a tendency in most cultures to rate oneself at the high end of a scale on a socially desirable quality, as well as to rate oneself in reference to one's own group. These tendencies not only bias results, but obstruct accurate comparisons across participants in a program, or across programs and cultures, and across time. Using reports by others along with self-reports, and focusing items on actual observable behaviors rather than endorsements of statements, may produce more objective results (Blades et al., 2012; Center for the Economics of Human Development, 2015). For example, the Flourishing Children Project's Goal Orientation scale includes the following behavioral item, "How often do you make plans to achieve your goals?" on a frequency scale from "none of the time" to "all of the time."

In addition, anchoring vignettes and situational judgment tests have been successful in reducing these biases and increasing validity and reliability, but require a more sophisticated and costly administration and analysis process, and situational judgment tests require a high level of literacy of respondents. They have not yet been validated to detect change over time.

Response scales: Response scales are often overlooked in reviews of instruments, but they are critical in determining the sensitivity of items to detect differences between program participants and within participants over time. Most of the instruments reviewed use simple Likert scales, which are good for identifying differences in general tendencies between individuals, but finer grain response scales are needed. Specifically, improved response scales could address the tendency toward an upward bias in self-report, by capturing variation at the upper end of scales to differentiate between youth who excel at a skill and those who are just above average (Lippman et al., 2014). Making such distinctions could establish thresholds that could help answer the question of how much of a skill is enough to affect an outcome. Finer grained responses at the upper end also allow for the detection of growth over time within an individual, due to a "ceiling" effect. If a youth rates highly at the start of a program, there is no room on the scale to detect growth. Measuring frequencies of behaviors, when possible, is more objective than the degree of endorsement by the youth of a skill, and can be used in reports by others as well (Lippman et al., 2014). When youth reports are triangulated with measures by others for more objectivity, it raises the additional challenge of making sure that both youth and adults or "other" reporters share the same concept/understanding of the skill, which is, of course, essential to model and develop the skill among youth.

Developmental appropriateness: There are differences in how skills manifest as youth age. The age span from 12–29 is large and encompasses huge differences in development, including cognitive processing, identity formation, emotional regulation and executive function, social contexts, life experiences, and academic, technical, physical, and practical skills, to name a few. Items need to be used, adapted, or developed that are appropriate for specific age groups and that reflect the youth's understanding of a skill and how it is demonstrated across contexts and relationships, such as school, work, with peers, or family members. Most measures found were for adolescents ages 15–19 rather than early adolescents or young adults, and so will need to be adapted or developed to suit all age groups of interest.

Measuring change over time: Research is needed on how to reliably measure change in soft skills at the individual level over time. This is needed specifically for program evaluations that

seek to determine whether a program has been successful in improving individual skills, but few measures have been validated for that use. Some assessment developers warn against using their measure for such purposes. Programs can succeed in educating youth and raising awareness about what is involved in a skill, and in giving youth practice using a skill, yet scores can decline in the program as a result of youth developing a more accurate self-perception of their skills in relation to others and to their own potential. The use of frequencies of behaviors along with reports by others may help to more accurately measure improvement.

Validation of instruments for program evaluation purposes: Many current tools in the inventory can be used for formative assessment—to inform youth so they can improve; and for program implementation purposes—to improve a program, but few were found that have been validated for program evaluation purposes. Specifically, the field needs tools that are sensitive to program interventions of short duration and that will detect change over time either at the individual or group level, depending on the evaluation design, and link performance on each skill to youth outcomes in order to discern how best to improve skills and improve youth outcomes.

Recommendations

An investment in tool development is recommended to provide the field with an improved measure of youth soft skills that is tailored to the needs of diverse international youth development programs.

- A soft skill assessment should be developed that draws from the universe of existing tools, is designed specifically for program use, and is appropriate for the age groups of interest. Adaptation might focus first on the high-scoring tools, supplementing as necessary with other relevant items or scales to adequately measure each skill independently, and include age and culturally appropriate language that can be ascertained through cognitive interviews.
- Such a tool should measure at least the three key cross-cutting skills (positive self-concept, self-control, and higher order thinking skills), using common terminology and definitions developed for this project that enjoy the strongest evidence across the fields of workforce development, violence prevention, and sexual and reproductive health. Preferably, a tool should also include additional skills that enjoy strong support for one or multiple outcome areas: communication, social skills, empathy, goal orientation, positive attitude, and responsibility (see the report, “Key Soft Skills for Cross-Sectoral Youth Outcomes” (Gates et al., 2016)).
- The instrument should be short and easy to administer, translated into languages needed for programs in Latin America, Africa, Middle East, and Asia, and the data resulting from assessments should be easy to analyze and report out.
- The measure should incorporate multiple methods to mitigate the shortcomings of self-report. This might include accompanying self-report scales with an observer report method such as program checklists and/or performance tasks, or at least a report from another person, preferably a program staff member. The items should measure frequencies of behaviors that can be reported on by the youth as well as others, which is more objective than endorsing statements. This will involve developing and testing new

response scales that accurately report upon and discriminate frequencies of behaviors, particularly at the upper end of the scale.

- Given the need for international adaptation, the instrument should be developed and pilot tested in multiple international program contexts and should preferably be validated for measuring change over time before being used to evaluate program contributions to soft skill development.

This investment would build upon investments in research on common skills and measures to date, enabling consistency in skill definition and measurement, and, once used by programs throughout the world, comparability across programs and evaluations, building the evidence in the field. An immediate benefit to programs would be provided by helping them target assessment and measurement efforts on the most important skills in a cost-effective manner. The long-term benefit is learning what works to improve youth skills in different contexts throughout the world and how that relates to youth outcomes across sectors.

2. Acknowledgements

We would like to thank the numerous experts involved in this study for their insightful contributions. In particular, the skill measurement experts and practitioners whom we interviewed were instrumental in contributing technical insights from their own experience, and input on and access to soft skills measurement tools. The full list of experts interviewed can be found in Appendix B. In addition, the report and inventory were greatly strengthened by the thorough review and thoughtful comments provided by external experts. Those included Koffi Assouan, MasterCard Foundation; Laurence Dessein, USAID; Clare Ignatowski, Independent Consultant; Richard Lerner, Tufts University; Karen Moore, MasterCard Foundation; Lee Nordstrum, RTI; Rebecca Pagel, EDC; and Rich Roberts, ProExam. The authors would also especially like to thank Elizabeth Berard, Cate Lane and Nancy Taggart of USAID for their support and technical guidance and reviews throughout this initiative.

In addition, the authors would like to recognize the following FHI 360 team members for their contributions. Ania Chaluda and Sara Babb reviewed measurement tools and relevant literature, and contributed to the design and development of the inventory. Elebthel Gebrehiwot, Alyson Mathews, and Clare Ofodile provided research support, identifying tools and relevant literature. Andrew Fine formatted the document and provided essential operational support. Kaaren Christopherson edited the report. Graphic design services were provided by the FHI 360 DesignLab. Invaluable technical reviews were provided by Kristin Brady and Michael Tetelman.

3. Introduction and Purpose

Soft skills are key for youth to succeed across multiple areas of their lives, including at school, at work, and in the larger community. Evidence demonstrates that soft skills foster a number of tangible health, well-being, relationship, education, and workforce-related benefits (Lippman et al., 2015; Deming 2015; Almlund et al., 2011; Heckman et al., 2006; Carneiro et al., 2007). Soft skills refer to a broad set of skills, behaviors, and personal qualities that enable people to effectively navigate their environment, relate well with others, perform well, and achieve their goals (Lippman et al., 2015).[1] As this evidence base on the importance of soft skills for fostering positive youth outcomes has grown, international youth development programs have increasingly focused on interventions that develop soft skills. This growth in soft skills-focused interventions has resulted in an urgent need among youth development programs for measures that can reliably assess key soft skills at an individual level over time, within a program implementation context.

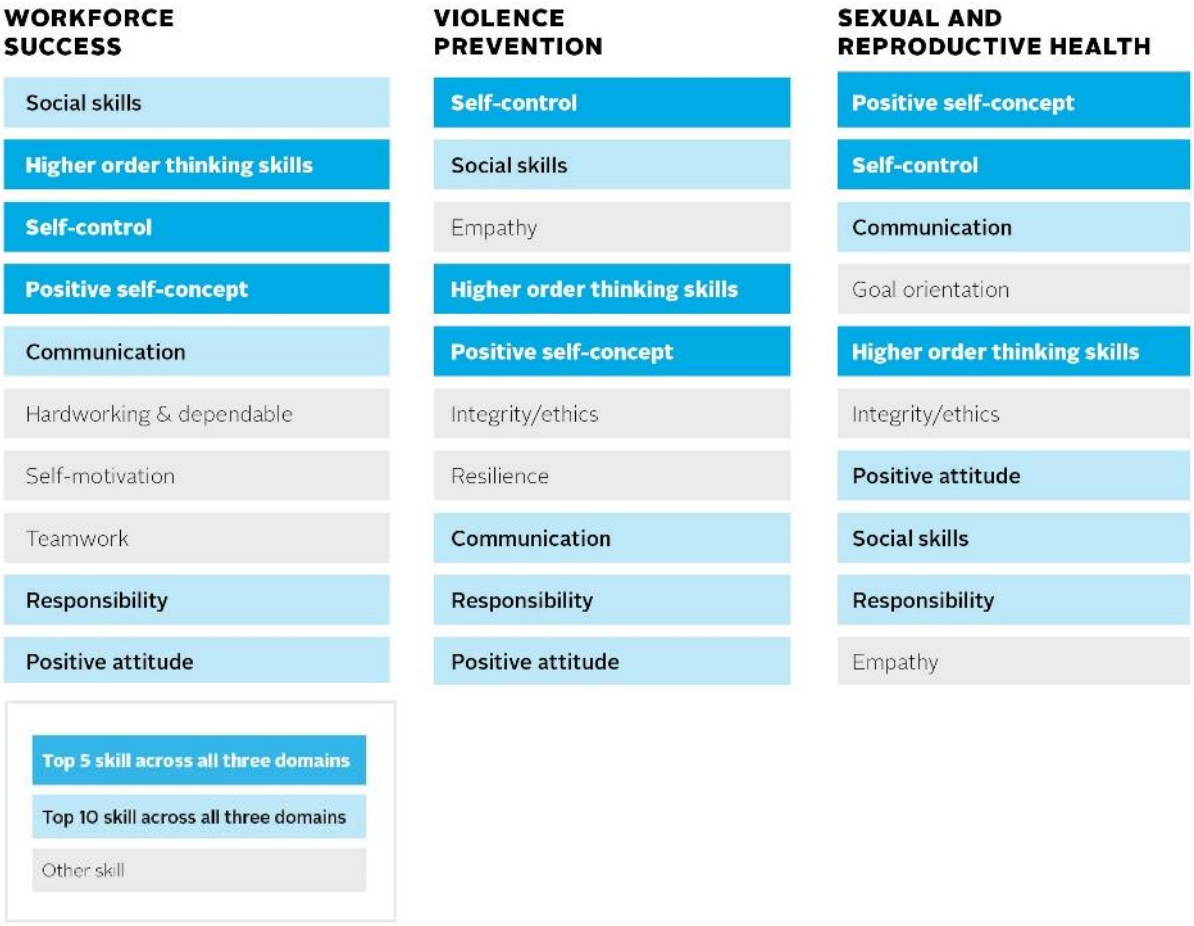
In efforts to advance research in this area and inform a cross-sectoral approach to programming, USAID has funded a series of studies on soft skills for youth development that focus on identifying the most important soft skills for key youth outcomes and on analyzing instruments to measure those skills (see Lippman et al., 2015 and Gates et al., 2016). In 2015, USAID published "Key 'Soft Skills' that Foster Youth Workforce Success," which identified the soft skills most critical to youth workforce success (Lippman et al., 2015). Building on that evidence base, USAID's YouthPower Action initiative conducted an extensive literature review to identify a common set of key soft skills that can help achieve positive outcomes across three different areas: workforce development, violence prevention, and Sexual and Reproductive Health (SRH). The report identifies seven skills that enjoy strong and wide-ranging support while being developmentally appropriate and malleable during ages 12–29.

Figure 1. Key Skills for Cross-Sectoral Youth Development: Top Supported Skills Across Fields

[1] Many terms have arisen from different domains to refer to similar sets of skills, including life skills, non-cognitive skills, and social-emotional skills. See pp. 238-239 of Duckwork and Yeager (2015) for a helpful discussion of skills terminology.

These are: higher order thinking, social skills, communication, self-control, positive self-concept, empathy, and goal orientation (see Figure 1). In addition to these seven skills, the skills responsibility and positive attitude received support across all three outcome areas, although to a lesser degree than those discussed above
(see Table 1).

Table 1. Most Supported Skills in the Literature within the Domains of Workforce Success, Violence Prevention, and Sexual and Reproductive Health



International youth development programs that work in this area need to be able to accurately measure soft skills among their youth beneficiaries in order to assess participants, improve programming, know whether interventions are improving the skills, and whether skill acquisition has an impact on outcomes. This measurement is needed to inform decision making about program design, implementation, and funding. Although there may be consensus that soft skills are important, however, there is less clarity on how to measure them. One expert interviewed for this research referred to the soft skills measurement landscape as the “Wild West,” while a Rand report on “Measuring 21st Century Competencies” describes “a dizzying array of options” (Soland et al., 2013, p. 9).

This report attempts to bring clarity to this field by inventorying measurement tools using objective criteria, building upon the evidence-based skills developed through the USAID investments described above. Our review focuses on the skills identified in the two literature reviews (Lippman et al., 2015 and Gates et al., 2016) previously mentioned as enjoying strong and wide-ranging support across multiple outcomes, specifically: higher order thinking, social skills, communication, self-control, positive self-concept, empathy, goal orientation, responsibility, and positive attitude. In addition, the review and inventory included other skills that were among the top 10 most supported skills of the workforce readiness report: hardworking and dependable, teamwork, and self-motivation.

After presenting general findings from the inventory of measures, this report establishes criteria for quality of measures of soft skills for youth development programs, and then reviews each tool based upon those criteria. A set of tools are then described that measure the top three skills and that are found to be of high quality according to the criteria. These tools may be promising starting points for programs searching for measures that are available now. Finally, the report identifies some of the challenges in measuring soft skills, and makes recommendations that would help to move the field of soft skill measurement forward to better serve the needs of youth development programs.

4. Research Landscape

State of the Field

Measurement tools may be used for a number of different purposes for international youth development programs, ranging from low to high stakes, including: 1) formative assessments, to inform program participants of their progress; 2) implementation, to provide programs with information for the purpose of better implementing their programs; 3) summative or descriptive, to describe or monitor the progress of youth at the group level within a program; and 4) evaluative, to evaluate the effectiveness of a program in developing skills or having an impact on specific outcomes through skills development. Measures may differ by their use in form, content, the nature in which scores are reported, and the level of standards applied. Multiple types of measurement tools exist, and there are a number of different ways to organize them (for examples, see ETS, 2012; Duckworth and Yeager, 2015; Kyllonen, 2015; Soland et al., 2013; and Stecher and Hamilton, 2014). We propose the following organization: 1) self-reports and self-rating; 2) reports and ratings by others; 3) performance assessments and simulations; and 4) mixed methods measures. In this section, we present definitions and a discussion of each of these types.

Self-reports and Self-ratings

Self-reports and self-ratings are the most commonly used types of soft skills measures because they are inexpensive, easy to use, and potentially reliable (Duckworth and Yeager, 2015). A common type of self-report is a questionnaire that asks youth to rate themselves on a Likert scale. This method is easier to score than an open-ended questionnaire or other types of multiple choice options, but it also presents several potential sources of error. These include reference bias, whereby frames of reference differ by individual according to their social group norms, and social desirability bias or faking, whereby individuals provide answers that they perceive to be “desirable” but are not accurate. (See page 17 for a more in-depth discussion of these sources of error.)

Several innovative approaches have been proposed to address these potential sources of error. These include anchoring vignettes, forced choice methods, and situational judgment tests. It is important to point out, however, that these methods are associated with their own trade-offs. They can be more complicated to administer and analyze, as compared to other types of self-reports, and situational judgment tests often have higher literacy requirements for respondents.

Anchoring vignettes present hypothetical situations and people that illustrate various skill levels, followed by a series of response options, one of which is correct. The respondent is asked to rate the vignettes on the same scale used for a self-report, which is administered at the same time. The respondent’s self-assessments are then compared to the respondent’s assessments of the hypothetical people described in the vignette(s). The self-reported response is then recoded to indicate whether it was lower than the respondent’s lowest rated vignette, at the level of the rated vignettes, or above the highest rated vignette, and this new score is analyzed

(Kyllonen). Kyllonen and Bertling (2013) have shown that anchoring vignettes can, in fact, help address response bias problems and increase cross-country score comparability. Likewise, results from the Anchored BFI demonstrate how incorporating anchoring vignettes and situational judgment tests (see discussion directly below) can improve cross-cultural comparability as well as the tool's predictive validity (Pagel et al., 2016).

Forced choice methods represent an additional innovation that can help to address the problem of faking in self-reports. As Kyllonen describes, from a respondent's perspective, "the best response is often to 'strongly agree,' with any statement that reflects a quality that an employer or school might value" (p. 201, 2015). Forced choice methods present youth with two or more options and asks them to choose the one that best describes them on a particular construct. The choices are intended to make it less clear what the "right" answer is, thereby reducing the likelihood of faking. Then, that choice is used to adjust other responses on the questionnaire. Evidence shows that, like anchoring vignettes, forced choice methods may increase cross-country validity and predictive validity compared with traditional self-reports (Bartram, 2013; Salgado and Tauriz, 2014).

Situational judgment tests (SJTs) can also help to address faking. SJTs present respondents with a scenario, typically through reading text that is longer than an anchoring vignette, meant to test their mindset or judgment related to a specific skill and asking them to respond, typically with a series of choices (although the format could theoretically be open-ended). Scenarios are frequently developed by asking individuals to describe a "critical incident" associated with a particular soft skill and then collecting various responses to the event that serve as alternatives that, while good, might not achieve the skill-related goal.

Other types of self-reports can include data gathered through personal essays/statements and biographical data, such as information on youth's extracurricular activities and accomplishments. Because biographical data may be gathered without students' knowing that it will be used to assess soft skills in particular, it may help to address bias. Methods for scoring this type of data are less well-established, however.

Reports and ratings by others

Ratings by others might be assessed through simple questionnaires or through more complex methods such as observational assessments. Evidence indicates that, on average, ratings by others have more predictive validity for educational and job success than self-ratings (Connelly and Ones, 2010; Oh, Wang and Mount, 2011, cited in Kyllonen, 2015). Kyllonen (2015) points out that, theoretically, anchoring vignettes or forced choice methods could be used to rate

others, but they have not traditionally been used this way since the problems that these methods aim to fix are largely addressed through using reports by others.[2]

In addition to questionnaire methods, reports/ratings by others can also take the form of observing youth behaviors that demonstrate a soft skill. Short "observation checklists" can facilitate rating these behaviors, which might include collaborative group work or presentations, as they are taking place. Records of the behavioral observation can also vary to include video, photographs, audio recordings, or notes (ETS, 2012).

It is important for raters to know well the youth that they are rating (Duckworth and Yeager, 2015; Kyllonen, 2015). Teachers may be able to provide this perspective; they also have the benefit of being able to compare youth with many other same-age youth. It may be the case that raters, such as program staff or teachers, only observe youth behaviors in one particular context, however, and miss the nuance that other adults, like parents, see across other contexts. Further, observers may misinterpret youths' behaviors, or their judgment might be clouded by their overall assessment of the youth (Duckworth and Yeager, 2015).

Performance assessments and simulations

Performance assessments and simulations are methods that require respondents to perform tasks that mimic real-life activities. Duckworth and Yeager (2015) describe the performance task method as "a situation that has been carefully designed to elicit meaningful differences in behavior of a certain kind." The most well-known performance assessment is the "marshmallow test" of delayed gratification (see Michel, 2014).

Although task-based measures may help to reduce measurement error that can occur through faking and the biases that may arise through self-reports and reports by others, they also cost more and take more time to administer. This method can also be plagued by "inconsistent scores across raters, tasks, and even a student's own performance on the same task repeated at different times." For this reason, performance assessments need to take place in highly controlled conditions.

Technology and game-based assessments have been introduced more recently to help address some of the above issues. Although they require an initial investment in required technology, they may reduce the cost of physical materials and time investments over time. They can also create a highly controlled environment and repeatable scenarios. The Knack app, for example, consists of three games that are designed to be psychometric assessments. Users download the app from a portal such as Google Play and create an account. The app then prompts the user to download any of the three games and provides guidance on how to play the game. It also provides explanations of digital badges a player gains by playing the game repeatedly.

[2] Kyllonen describes the behaviorally anchored rating scale (BARS), another popular method for collecting others' ratings. Although this scale is primarily used for cognitive testing, it provides promising direction for soft skills measures. The scale contains "behavioral anchors" collected through critical incidents that help to "provide additional meaning for the score points…." For example, a BARS measure for "analytical reasoning" includes anchors along a scale from 0–5; a score of 4.6 for "analytical reasoning" is described as "extracts the essence of complex issues and doctrines quickly and accurately."

Finally, in some job markets, the app can connect players to relevant job opportunities based on the digital skills badges they earn.

Finally, direct assessments, whereby a measure is embedded within an assessment (such as a questionnaire), can be considered a type of performance assessment. This method might assess problem solving directly by posing problems for the respondent to solve, rather than relying upon the respondent's or other's reports. The PISA (Program for International Student Assessment) Problem-Solving computer test is an example of this.

Assessments may also include a mix of the methods described above; triangulating methods can help to address sources of measurement error. This is usually done through combining a youth self-report with a parent-, teacher-, or other observer (e.g., program facilitator) report. For example, the Buck Institute for Education Presentation Rubric consists of both a student self-assessment and a guided review that teachers can use to assess students' performance. The Jovenes Constructores Competencies Self-Evaluation consists of both a youth self-report and an observer report that is completed by someone who has observed the youth's behavior throughout the program.

Siloed Development of Measurement Tools

Although there is general agreement about the promises and pitfalls of the different methods for measuring soft skills, there is less agreement on how skills should be conceptualized and grouped together. Measures of soft skills have been developed out of different sectors and traditions, including education, youth development, workforce development, psychology, and public health, as well as by different types of stakeholders, including practitioners, funders, policymakers, and researchers. A key informant interviewed in a RAND research study on measuring skills explained the problem: "People often complain about this Tower of Babel and the different labels. Each little subdiscipline has its own traditions, its own language, its own codes, which is part of the problem (quoted in Stecher and Hamilton, 2014p. 38).

This siloed development has resulted in conflicting approaches to soft skills measurement at different levels, including: 1) skill conceptualization (e.g., employability skills, non-cognitive skills, socio-emotional skills); 2) skill domain taxonomy (the way types of skills are grouped together according to underlying theory, such as "the Big Five personality factors"); and 3) skill identification (the way specific skills are named and defined, such as "teamwork" or "collaboration"). Different conceptualizations of skills as "non-cognitive," or "socio-emotional," though not exactly the same, are generally in agreement about core characteristics of these skills—for example, that they are relatively stable over time, responsive to intervention, and dependent on context for their expression—and that when "terms with similar meaning are grouped together, a substantial consensus emerges around which types of skills are considered most useful" (Duckworth and Yeager, 2015; Lippman et al., 2014, p. 13).

Duckworth and Yeager argue that "from a scientific perspective, agreement about the optimal terminology for the overarching category of interest may be less important than consensus about the specific attributes in question and, in particular, their definition and measurement" (2015, p. 239). The field is plagued with diverging definitions and measures of the same skill,

while at the same time, there are many instances where one measure is used to represent multiple skills that are conceptually and empirically different from each other. This paper and other YouthPower Action research attempts to move the field toward a consensus on skill terminology and definitions by using terminology previously proposed in "Key 'Soft Skills' that Foster Youth Workforce Success: Toward a Consensus Across Fields" (Lippman et al, 2015) and expanding it from a workforce focus to include terms in the fields of violence prevention, and sexual and reproductive health as discussed in "Key Soft Skills for Cross-Sectoral Youth Development" (Gates et al. 2016).

Methodological Challenges in Measuring Soft Skills in International Youth Development Programs

Determining Reliability, Validity, and Measurement Invariance

A lack of high-quality soft skills measures poses problems for programs intending to use measures to understand youth's improvement in soft skills. Ensuring technical quality means considering fundamental psychometric principles—namely reliability, validity, and measurement invariance—and balancing these with cost and ease of use (Soland et al., 2013). It is also important to recognize that more recent, complex testing methods, for example task-based tests or simulations, might introduce new sources of measurement error and may require new approaches for technical quality control (Soland et al., 2013).

Reliability refers to consistency. A tool can be considered reliable if the respondent, taking the test again under similar circumstances, receives the same results. Inconsistency can stem from measurement error, which might result from a variety of sources—internal sources such as poor correlation among question items or external sources such as disagreement on scoring among raters. Regardless of the source of the error, tests with low reliability will not provide useful information (Stecher and Hamilton, 2014).

Reliability must also be balanced with validity, which is arguably the most important technical consideration for measurement tools. Validity, in brief, is the extent to which a tool measures what it purports to measure. Test developers can use multiple sources of evidence to establish good validity, including: correlations of the measure with other related measures; evidence of the ability of the measure to predict intended outcomes; expert evaluations of the representativeness of the items; and interviews to determine whether the test elicits intended the responses (Stecher and Hamilton, 2014). Some of these sources of evidence are easier to capture than others. Stecher and Hamilton describe how to strategically assess validity: "Developers and users should clearly identify the purpose of a measure along with the inferences that it is intended to support; develop an argument linking these inferences to the types of evidence that support them (Kane, 2006); and devise a plan for gathering this evidence" (2014, p. 41).

Finally, measurement invariance must also be considered in assessing a tool's psychometric rigor. Measurement invariance refers to whether a tool performs similarly or differently across multiple groups (e.g., gender, age, ethnicity, region) and can be assessed empirically through item response theory analysis or confirmatory factor analysis. These analyses are advanced

and can be time-consuming, however; thus, this evidence, which could be richly informative for future tool development, is often lacking (Card, 2016, in press).

Self-reporting Methods

As described earlier, self-report and other-report questionnaires are the most commonly used methods for soft skills measurement, for well-established reasons. These methods, particularly self-reports, are uniquely vulnerable to measurement error, however. First, self-report and other-report methods assume that the user understands the question's intended meaning, which may not be true if the respondent has low literacy levels or misinterprets questionnaire items. Self- and other-reporting methods also require individuals to coherently recall and summarize information in order to make judgments, which may be colored by individuals' tendencies to see themselves or others as stable over time. Individuals must also use frames of reference to arrive at their judgments, and different youth will have different frames of reference. In other words, different youth will have different ideas about what being "very good" at a skill might mean, which will influence their responses (Soland et al., 2013; Duckworth and Yeager, 2014; Kyllonen, 2015). Individuals may also adjust their responses to be "socially desirable," or to meet external standards that they perceive.

Measuring Change over Time

Using a soft skills measurement tool to measure change over time is particularly challenging. Soft skills measurement tools have been used extensively to measure difference in skills among populations at one point in time; however, methods to accurately measure change in those skills over time are still lacking. This problem is primarily due to reference bias among respondents, who may initially perceive their soft skills abilities to be high before they have a thorough understanding of the full continuum of abilities required of the skill. Catholic Relief Services found in a recent study that the soft skill scores of participants in their Jovenes Constructores program initially declined as they learned more about what the skill entailed, but then increased as they gained proficiency in it (Herman, 2016). Duckworth and Yeager (2015) remark that: "In fact, current data and theory suggest schools that promote personal qualities most ably—and raise the standards by which students and teachers at that school make comparative judgments—may show the lowest scores" (p. 244).

Other reasons for challenges in measuring change over time include faking due to high accountability pressure (Duckworth and Yeager, 2015) and a lack of differentiation at the higher end of scales that allow youth to demonstrate skill fluency (Lippman et al., 2014). Individuals' tendencies to rate themselves highly imply the need for nuanced differentiation at the upper end of scales in order to demonstrate any improvement (Lippman et al., 2014). Overall, it is important to triangulate methods and incorporate performance-based measures where possible (personal communication with Ana Maria Munoz-Boudet, July 27, 2016; personal communication with Patrick Kyllonen, May 13, 2016; Duckworth and Yeager, 2015).

Developing and Adapting Tools for Use across Cultures and Levels of Resources

Most soft skills measures have been developed, tested, and validated in developed countries. This can be problematic due to reference bias when measures are not "anchored" in an objective phenomenon. In other words, individuals will have different ideas of what a quality such as "self-motivation" means, and might respond differently to test items across contexts (Kautz and Heckman, 2014). As one expert interviewed for this project explained: "You could have an instrument that has been used in the U.S. and in [another] country … [but] that doesn't necessarily say that you could transfer that to an environment where there is a different religious and value perspective in the community."

Methods that anchor skills in an objective phenomenon can help in adapting tools across cultures. Anchoring vignettes may help to address this problem by providing respondents with a reference point for their judgments. Forced choice methods, by asking respondents to choose among various options, can also serve this purpose. Kyllonen and Bertling (2013) have illustrated how anchoring vignettes enhance within-country validity and cross-country score comparability of the PISA assessment. Although these methods have been successful at improving the cross-cultural and predictive validity of the Big Five personality factors inventory, they have not yet been shown to be reliable for testing change of a skill over time.

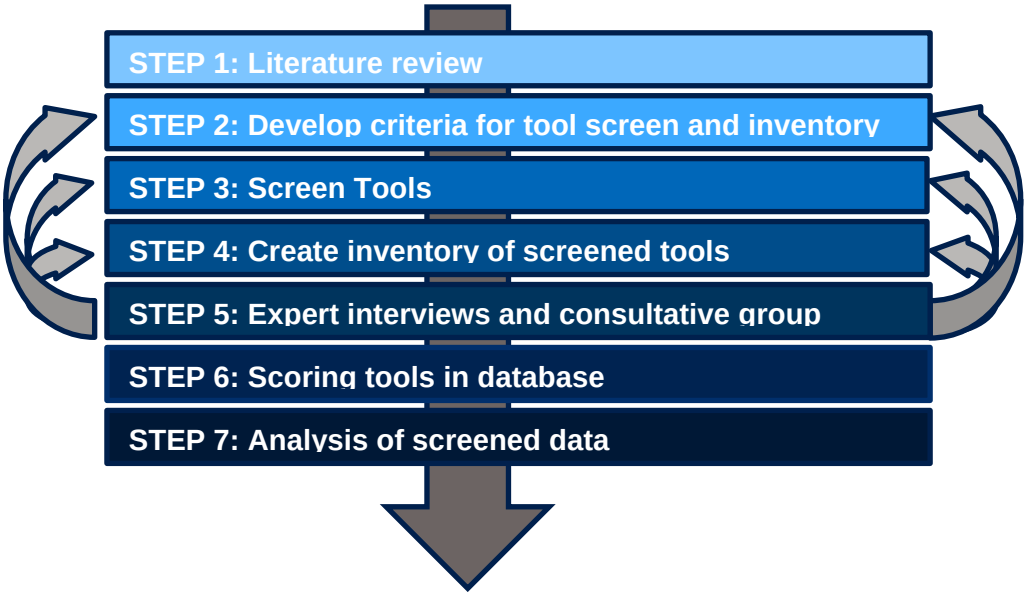
Several other contextual considerations are worth noting. In addition to understanding cultural perceptions of soft skills, it is also important to be mindful of a country's literacy levels and the extent to which the test-taking or administration relies on literacy. Where structural resources are lacking, it is important to consider the extent to which measures can adapt to multiple delivery mechanisms, from pen and paper to computer-based methods.

5. Methodology

USAID YouthPower Action has identified a set of key skills that foster positive outcomes across the domains of workforce development, violence prevention, and SRH. These include positive self-concept, self-control, higher order thinking skills, social skills, communication skills, goal orientation, empathy, responsibility, and positive attitude, as defined by the skills framework proposed by "Key 'Soft Skills' That Foster Youth Workforce Success" (Lippman et. al 2015) and "Key Soft Skills for Cross-Sectoral Youth Outcomes" (Gates et al., 2016). International youth development programs need tools that can measure the level of these skills among participants and ideally, the development of these skills during a program. Building on the findings of the two above-mentioned reports, this report reviews existing tools that may be relevant for program use, which have been compiled in a detailed inventory designed to be an informative resource for practitioners and researchers alike.

The measurement tool review process began by identifying all tools that might measure one or more key soft skills of interest for youth workforce, SRH, and violence prevention outcomes. Next, tools were evaluated using a standard set of review criteria developed, based on the review of existing tools, expert perspective, and input from practitioners. Criteria were applied across a two-stage review process. Tools were first screened for relevance to the key soft skills of interest, as well as age appropriateness and cost. Those that passed the initial screen were then reviewed in-depth for evidence of validity, reliability, outcomes of interest, ease of use, evidence of international use, and assessment type. Finally, the tools were divided into three groups based upon the degree to which they met all of the criteria: high (meeting 5–7 criteria), medium (meeting from three to fewer than five criteria), and low (meeting fewer than three criteria). The steps used to identify tools, screen, score, and analyze the results are summarized below in Figure 2.

Figure 2: Steps in Identification of Measurement Tools



Literature Review Search Strategy

The first stage of the measurement tool review began with identifying soft skills measurement tools. This process took place from February 2016 to October 2016 and resulted in identifying 244 tools. The process began with a review of more than 150 measurement tools that were documented as part of the research on "Key 'Soft Skills' That Foster Youth Workforce Success." Tools were also identified through the research on "Key Soft Skills for Cross-Sectoral Youth Outcomes" and reviewed. Next, the team searched key synthesis reports, such as "From Soft Skills to Hard Data: Measuring Youth Program Outcomes" (Forum for Youth Investment, 2014) and "Measuring Hard to Measure Student Competencies" (Stecher and Hamilton, 2014), as well as soft skill measurement databases that were identified by a senior staff member.

These steps were accompanied by interviews with measurement experts, public and private tool developers, and practitioners in the international youth development field (see Appendix B for a list of key informants interviewed), who recommended measurement tools and/or databases of tools. The team also requested relevant measurement tools from YouthPower Learning's Cross-Sectoral Skills Community of Practice via e-mail and in person on May 16, 2016. The initial search was restricted to English language tools; however, in the course of interviews one non-English tool was recommended by an expert and relevant documentation was translated and reviewed from the original Portuguese version.

Finally, a broad online search was conducted to supplement the specific strategies described above. The following academic databases and Google Scholar were searched: Pubmed, Popline, Global Health, Africa Wide Information, PsycInfo, Education Full Text, ERIC, and Social Work Abstracts. These searches were restricted to 1990 to 2016 and returned 26 additional papers on tools (that were not duplicates of those already reviewed).

Selection of Screening Criteria and Tool Review

Central to the tool screening process was the development of the criteria to screen and categorize identified relevant measurement tools. For efficiency, a two-stage process was developed. The first stage was designed to remove tools that did not address any of the key soft skills of interest, did not target the age range specified by this research effort (defined below), and/or had prohibitive costs associated with use of the tool (described below). (See Appendix C for a list of tools reviewed.) Tools that made it through the first screen were reviewed a second time in greater detail; information on the tool design, use, and psychometric properties, such as validity and reliability, was documented. This process resulted in an inventory of 74 tools.

The screening process was conducted by the research team; four team members systematically recorded information into two screening databases and discussed their results regularly with other team members. Tools were assessed in groups of 25 and discussed at intervals of 5 tools to ensure consistent and accurate coding over time. The tools were also cross-checked by a researcher who was not involved in the initial coding. This process helped to ensure coding reliability and allowed team members to discuss and adapt the criteria.

Characteristics and Criteria for Screening

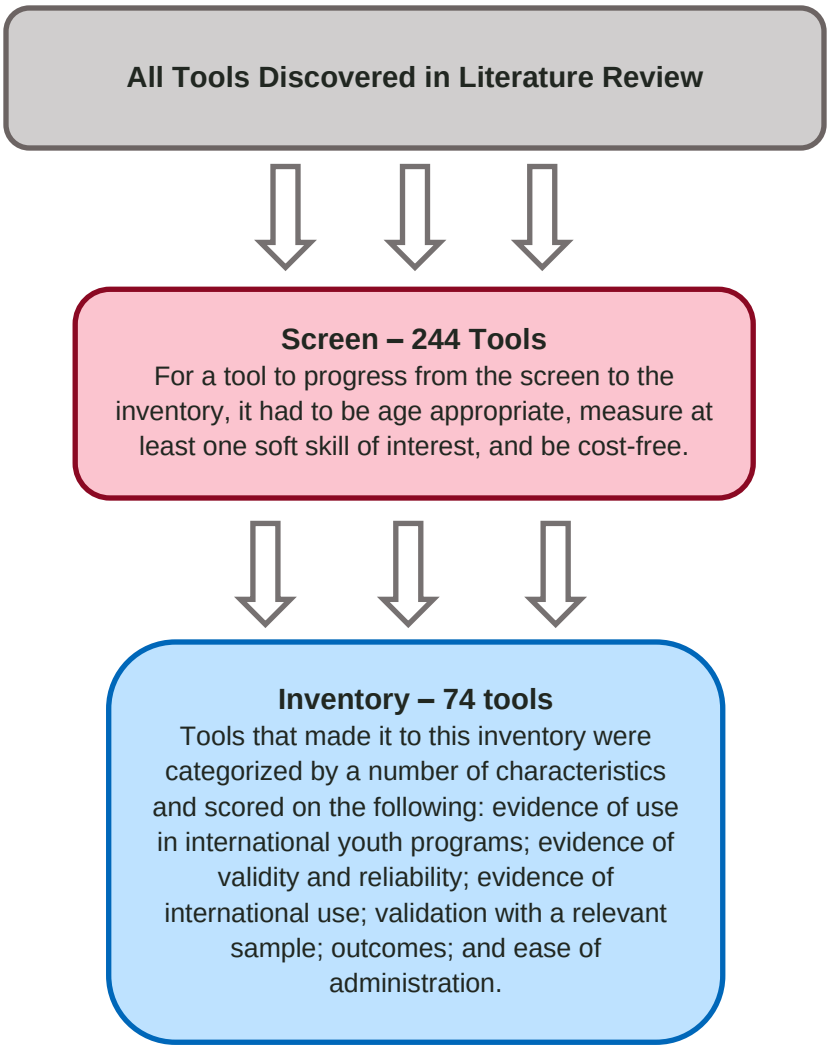
The initial set of screening criteria was developed by the project team, which included practitioners and researchers with expertise in youth development, education and workforce development, SRH, and evaluation science. These criteria were confirmed or adapted throughout the review process with external experts and practitioners who contributed recommendations on how to better refine the criteria. This feedback was gathered through 12 phone and in-person interviews. These interviews also helped to uncover additional tools and points of contact for accessing these tools.

Three of the four in the first set of screening criteria were assessed on a pass/fail basis, meaning they had to be met for a tool to advance to the next level of screening into the tool inventory, thus refining the list. The other criterion did not have to be met for the tool to move onto the next stage.

In the screen, in regard to the pass/fail criteria, a tool qualified for the inventory if 1) it measured at least one soft skill of interest; 2) there was an overlap between the age group for which the tool was created and the project's target population of 12- to 29-year-olds; 3) it was free (apart from rare exceptions). Contextual appropriateness was assessed from available documentation, but not used to determine whether a tool was going to be selected for the inventory due to its somewhat subjective nature and/or limited access to complete information about the use of a tool in developing country contexts. During this stage 244 tools were screened. The criteria applied in the first screening are shown in Table 2.

TABLE 2 SCREEN CRITERIA AND DEFINITIONS

Tool Content Validity (for soft skills identified) (if no, exclude)	The extent to which a tool measures any of the twelve key skills identified in both reviews of the literature.
Age Appropriateness (if no, exclude)	The target age group was defined as the age group for which the tool was created or administered. This activity specifically targeted tools that had been developed or used for youth ages 12–29 years.
Contextual Appropriateness—staff noted any relevant information, but lack of information did not preclude further consideration	Contextual appropriateness referred to the extent to which there was evidence that the tool had been used for or could be adapted to different contexts and diverse populations that represent the youth populations served in USAID programs. Here, adaptability refers to the language used in the tool and whether it is exclusive to any particular subgroup (e.g., gender, ethnicity, socio-economic status, urban/rural).
Cost of tool (if applicable—if yes, exclude)	The cost criterion indicated whether there were direct costs associated with the use of the tool. Tools with recurring fees for services and materials such as analysis, forms and booklets, or access were excluded, since they could be prohibitively expensive for a project. The criterion did not exclude tools had minor one-time fees that give the user unlimited use or if the tool developer was willing to provide special access to the tool, they were retained in the screen. However, this was encountered in only one case.



Inventory Characteristics

Seventy-four tools passed the first screen and proceeded to an in-depth review and categorization process. This process focused on criteria identified by our team, measurement experts, tool developers, and practitioners as relevant for practitioners interested in using a tool with USAID youth beneficiaries. No tools were excluded or removed at this stage. The ultimate purpose of the inventory is to describe each tool according to a set of key characteristics (see Table 3) to allow programs to compare across the tools and select tools that are most appropriate for their purposes. Some categories are strictly intended to provide information that might be of interest to programs, while a subset of the criteria, described on pages 25-26, were used to create a score evaluating a tool's quality and usefulness.

This review and categorization closely examined written information about the tool's psychometric properties and other characteristics indicating its appropriateness for international youth development programs. The tool characteristics noted in the inventory are shown in Table 3.

TABLE 3: INVENTORY CHARACTERISTICS AND DEFINITIONS[3]

Bibliographic information	Information on the author, citation information, and name of assessment
Type of assessment	Identification of type of assessment. Categories are 1) self-report/ratings; 2) reports/ratings by others (including "reports by others" questionnaires and observation checklists); 3) performance assessments and simulations (including direct assessments, performance assessments, and games); and 4) multiple assessment types.
Use	A narrative explanation of specifics about the purpose of the tool (evaluation of certain skills, prompting student self-reflection, etc.)
Age range	Specification on the range of ages of the population whose soft skills are measured by the tool.
Evidence of international use	Narrative evidence of use of tool in an international context.
Administration characteristics	Characteristics of the administration of the tool: time it takes to administer the tool; evidence of translation of the tool into languages other than English; training needed to implement the tool and/or analyze its results; and, when available, instructions on how to administer the tool (including information on location or size of participant population).
Key soft skills identified using common terms	The key soft skills of interest that a tool assesses using common terminology for this project, as coded directly from the instrument. These skills include any of the top 10 skills identified and defined in the report "Key 'Soft Skills' that Foster Youth Workforce Success," plus empathy and goal orientation, found to be important for predicting violence prevention and SRH outcomes in the report "Key Soft Skills for Cross-Sectoral Youth Outcomes."
All soft skills tested (using authors' terms)	All of the soft skills that a tool assesses using the tool developer's terms rather than the common key soft skills terminology (see above).
Outcome of interest tested for by tool	Key outcomes of interest that a tool has been used with, specifically four categories: workforce development, violence prevention, SRH, and other. The category "other" included important outcomes such as substance abuse or academic performance.
Number of questions for each construct	The number of questions for the construct of interest in the measurement tool. Modes of assessment, such as games, that rely more on tasks and actions will not have a count here.
Type of scale by which different responses are recorded	Description of the type of response scale the tool used, such as a Likert-type scale ranging, for example, from "strongly agree" to "strongly disagree" or from "never" to "always."
Sample population tested	Description of the sample population on which a tool is tested.

[3] No characteristics in this categorization were used to screen out a tool. They are only used for assessing a tool, providing information that will drive scoring, or to provide information to implementers.

TABLE 3: INVENTORY CHARACTERISTICS AND DEFINITIONS[3]

Bibliographic information	Information on the author, citation information, and name of assessment
Validity	Description of evidence that an instrument measures the construct it intends to measure, and its relationship to outcomes. Each construct is considered in relation to each outcome, and evidence of concurrent or predictive validity is recorded, when available. Types of validity that are recorded are construct, concurrent, predictive, and convergent/divergent validity.
Reliability	Description of whether an instrument measures the construct consistently across items and over time, and whether a scale measures a common underlying factor. Types of reliability that can be tested include internal consistency, test-retest reliability, etc.
Easy to adapt to international projects	Review as to whether the questions in the tool are broad enough to be applicable across different contexts. Here "adaptable" was defined as a tool that did not include language in its items that was exclusive to a geography, sex, ethnicity, or socio-economic stratum.
Cost	Discussion of cost, if any exceptions are made, otherwise notation of no cost.
Fairness to different groups, on ethnicity, gender, etc.	Review as to whether the measure has been tested with boys/girls, various age and ethnic groups, etc. This contains any data on measurement invariance found in documentation on a tool. Specifically, "fairness"' refers to whether the questions discriminate against one group (for example, only focusing on activities in which boys or girls are more likely to be involved). This is based on the general interpretation of the researcher, unless otherwise noted by documentation, noting instances of tool bias.
Number of skills of interest measured	A numeric value of the number of skills tested by the tool that fit under those targeted by this effort.

In addition to looking more closely at tool characteristics, the inventory focused on the quality of the instruments by examining their reliability and validity. In most cases, reliability and validity information was provided by the developer of the tool, but occasionally it was found elsewhere especially if a tool had been widely used and tested by other researchers and practitioners. As much as possible, any provided coefficients of reliability and validity were listed, and every specific outcome against which a tool was tested was noted. Furthermore, the researchers continued to investigate the extent to which a tool could be easily adapted for use in other countries, especially those where USAID is present, by looking at the samples on which the tool was tested, whether the tool was tested internationally, and whether the tool was fair to all groups, especially focusing on gender. Finally, the researchers matched the soft skills each tool measured to the soft skills that are the focus of this project, by carefully coding the skills based upon the common terminology and definitions of the skills developed in the annexes of the "Key 'Soft Skills' That Foster Youth Workforce Success: Toward a Consensus Across Fields" paper.

Scoring Measurement Tools

The final stage of analysis of the tools in the database involved summing the scores assigned to key criteria from the inventory (see Table 4). The purpose of the scoring is to provide numeric values that allow users to assess the extent to which a tool meets the criteria for being most useful, based on a practitioner's particular needs.

The research team reviewed each tool and scored it according to details captured in the database. A researcher then reviewed the tools for the inventory a second time, focusing specifically on issues of reliability and validity, and outcomes of interest. Any inconsistencies or required changes would prompt a review of existing literature on a tool, outreach to the tool developer, or discussion among the research team.

Table 4 describes the scoring criteria, how they were defined, and the range of scores a tool can receive.

TABLE 4: SCORE CRITERIA DEFINITIONS AND RANGE

Score Criteria	Definition	Score range
Evidence of use in international youth programs	This indicates whether or not the tool is suitable for use in programs internationally, based on its current use abroad.	0–1
Evidence of validity	This score indicates the tool's level of validity evidence. Where a tool has not been tested for validity or no information exists to confirm that it has been tested, it receives a score of 0. Where tool documentation indicates that validity testing has been done, but no empirical evidence (data) could be found, the tool receives a score of 0.5. Finally, where tool documentation demonstrates empirical evidence (data) that a tool has a validity coefficient of 0.3 or higher, it receives a score of 1. In Table 6, a score of 0.5 is indicated as "yes" and a score of 1 is indicated as "yes*."	0–1
Outcomes of interest	This is a four-part category to denote the three main outcomes of interest for this measure database (workforce development, SRH, violence prevention), and a fourth outcome ("other") that denotes other categories that are of interest to practitioners, but fall outside those three specified in the study design. In this category, a tool receives a 0 if no outcomes of interest are tested with the tool, or if no information on outcomes of interest were found. In a few cases, documentation suggested that a tool might be used to test for a particular outcome, but no evidence that it had been used to test for that outcome was found; in these cases, the information on outcomes was recorded in the inventory, but was NOT included in the score. A tool receives a 1 if an outcome of interest is tested. A tool receives a 1 with an * if any outcome tested by the tool falls into one of the three main outcomes of interest.	0–1

TABLE 4: SCORE CRITERIA DEFINITIONS AND RANGE

Score Criteria	Definition	Score range
Evidence of reliability	This score indicates the tool's level of reliability evidence. Where a tool has not been tested for reliability or no information exists to confirm that it has been tested, it receives a score of 0. Where tool documentation indicates that reliability testing has been done, but no empirical evidence (data) could be found, the tool receives a score of 0.5. Finally, where tool documentation demonstrates empirical evidence (data) that a tool has an alpha reliability coefficient of 0.7 or higher, it receives a score of 1. In Table 6, a score of 0.5 is indicated as "yes" and a score of 1 is indicated as "yes*." In cases of inter-rater reliability, a different threshold of 0.41 was used.	0–1
Evidence of international use	This indicates whether a tool has been tested in an international context. If a tool was used in only one high-income country, it received a score of 0. If it was used in at least two high-income countries (but no low- or middle-income countries), it received a score of .5. If a tool was used in at least one developing country, then it received a score of 1.	0–1
Relevant validation sample tested	This indicates that the tool was tested on the youth population targeted by this research (ages 12–29). A tool receives a 0 if the tool has not been tested with a relevant validation sample, or if no information is found. At a minimum a valid sample requires that youth be in the age range of interest. This ideally also implies youth populations in low- and middle-income countries, but a tool is not penalized if a tool is not tested outside the United States as the evidence of international use category already captures this characteristic.	0–1
Ease of administration	This comprises a three-part score. The three elements are: “no trained personnel required,” “available in languages other than English,” “short length of time to administer.” The short length of time is factored based on the time it takes to answer questions for each construct, not total time taking the test. The answers to each in sum equal the ease with which a person or program might use or adapt the tool to their own efforts. Each answer counts for one-third (1/3) of a point.	0–1

The scoring system provides equal weighting (i.e., a maximum score of 1) among criteria.[4] Aggregate scores were generated for each tool and ranged between 0 and 7. Although most scores are 0 to 1, one, ease of administration, is divided into three parts. The tools were then divided into three groups based upon the degree to which they met all of the criteria: high

[4] We recognize, however, that certain criteria (e.g. validity, ease of use) may be more important for specific users. Users may wish to apply a different scoring system than the one used in this inventory by prioritizing certain criteria over others.

(meeting five to seven criteria), medium (meeting from three to fewer than five criteria), and low (meeting fewer than three criteria).

It is important to note that the use of the score is not meant to denote a value judgment per se, since tools will be valuable to users based on their own needs and contexts. Specifically, these terms refer to the number of criteria met, providing an indicator of the level of availability and quality of evidence based on those criteria. For example, a top scoring "high" tool may have the following characteristics:

- Can be used internationally since there is evidence that the tool has been adapted to other country contexts or used abroad
- Easy to use, as seen by a positive rank on two of the three score elements
- Has been validated and shown to be reliable, and is used to measure outcomes that are relevant to key USAID youth development objectives
- Questions for each construct are short, meaning less likelihood for survey fatigue with youth participants

At the other end, a "low" ranked tool is still one that could be useful for an international youth program, but has shortcomings in relation to more highly ranked tools. For example, a low-ranked tool:

- May lack evidence of use outside of the United States.
- May be less easy to use.
- May not have been validated or tested for reliability.
- May require trained personnel or may take more time to administer.

In addition to the key criteria described above, other characteristics about a tool's performance or use were recorded to provide more information for analysis, but they were not factored into the score of the tool. This additional information about the tools was used in describing the tools and exploring trends across the tools in the database. The numbers they were assigned were not values, but used for assigning a metric that could be used in data analysis. For example, we describe the type of assessment employed by the tools numbers one to seven, (based on the seven types of assessments), whether tools were validated for use in measuring change in behavior over time (we noted if no information was provided on use in measuring change over time, if it was validated to measure change over time, and if it explicitly was invalid for use in measuring change over time). The number of soft skills of interest that tools measured were also scored, with 1 denoting a relevant skill a tool tested and 0 for its absence.

Limitations of Methodological Approach

It is important to note a few limitations of the methodology. First, although this project has identified many key soft skills measurement tools, it should not be considered a comprehensive list. The intention was to identify those tools that are most commonly used in the field and met the inclusion criteria described on page 23 In addition, the screening process excluded tools whose use would require a fee for use and/or analysis (i.e., beyond a one-time fee).

Second, this identification and screening of tools represents those tools that were available as of 2016. Tools may have been excluded due to their incomplete nature, or because they are still undergoing validity and reliability testing. In addition, as the search for relevant soft skills tools was extended, the focus on particular tools was further refined, to better target the scope of work intended under this project. In addition, some tools designed for very specific subject matter are excluded, since their adaptability to a broader context could be difficult.

Third, this collection of tools forms a disjointed landscape, in which terminology and approaches vary across contexts, industries, and cultures. Therefore, comparisons of tools can be difficult, which underlines the importance of adapting tools to specific contexts. Although the focus of this project is on tools that would be appropriate to contexts and countries in which USAID works, many tools we identified targeted U.S. or other Western educational contexts. Some of these tools require extensive use of computers and technology or sophisticated training to use or analyze data collected by the tool. Other tools are designed for use by institutions with high levels of resources or expertise for data collection. In both cases this makes tools potentially harder to adapt to low- and middle-income country contexts.

Finally, given the breadth of contexts in which soft skills are measured, it should be apparent that no one tool is capable of meeting all the measurement needs of youth development programs. The purpose of the list of tools that have been categorized in the inventory is to describe the breadth and depth that current tools reach both in the soft skills they measure and in the potential uses they may serve. The findings should not be used to definitively mark one tool as more useful or “better” than others; instead the scores are meant to describe differences among the tools with respect to measuring key cross-cutting soft skills for youth. Their usefulness will vary according to the needs of each program, including the skills that are the focus of the programs, the age group participating, and the purpose for which the tool will be used.

6. Analysis of Findings

In this section, we present our analysis of the 74 tools in three parts. First, we review tools that measure the key soft skills (as defined in the Methodology) to assess how widely those skills are measured by existing instruments and where there may be gaps. Feedback from interviews with practitioners and experts in the field has indicated particular interest in analyzing the universe of measures from the perspective of specific skills of interest for particular projects or contexts.

Second, we provide a brief descriptive overview of the entire universe of 74 tools reviewed in depth, according to review criteria. The overview covers areas that include the ease of use of tools, and challenges around tool appropriateness for measuring change over time, demography, and geography.

Third, we identify a shortlist of high-scoring tools that measure key cross-cutting skills for the purpose of use in international youth programs. We focus on measures available for the three key skills that receive the strongest support across workforce development, violence prevention, and SRH (positive self-concept, self-control, and higher order thinking). These high-scoring tools meet most (five to seven) of the criteria for usefulness for international youth development programs. We review the strengths and weaknesses of this body of tools as well as of each tool individually.

Readers should refer to the full inventory of measures for detailed information on each tool, including scoring results. The inventory will be posted online at www.youthpower.org.

Summary of Measures Available for Each Key Skill

Consultation with experts indicated that funders, implementers, and researchers want to understand which tools are available for specific skills. In response, this section presents the number of tools available, by skill, as well as how they scored against the review criteria. We also review the types of tools available to measure each skill (self-report, observer checklists, etc.).[5]

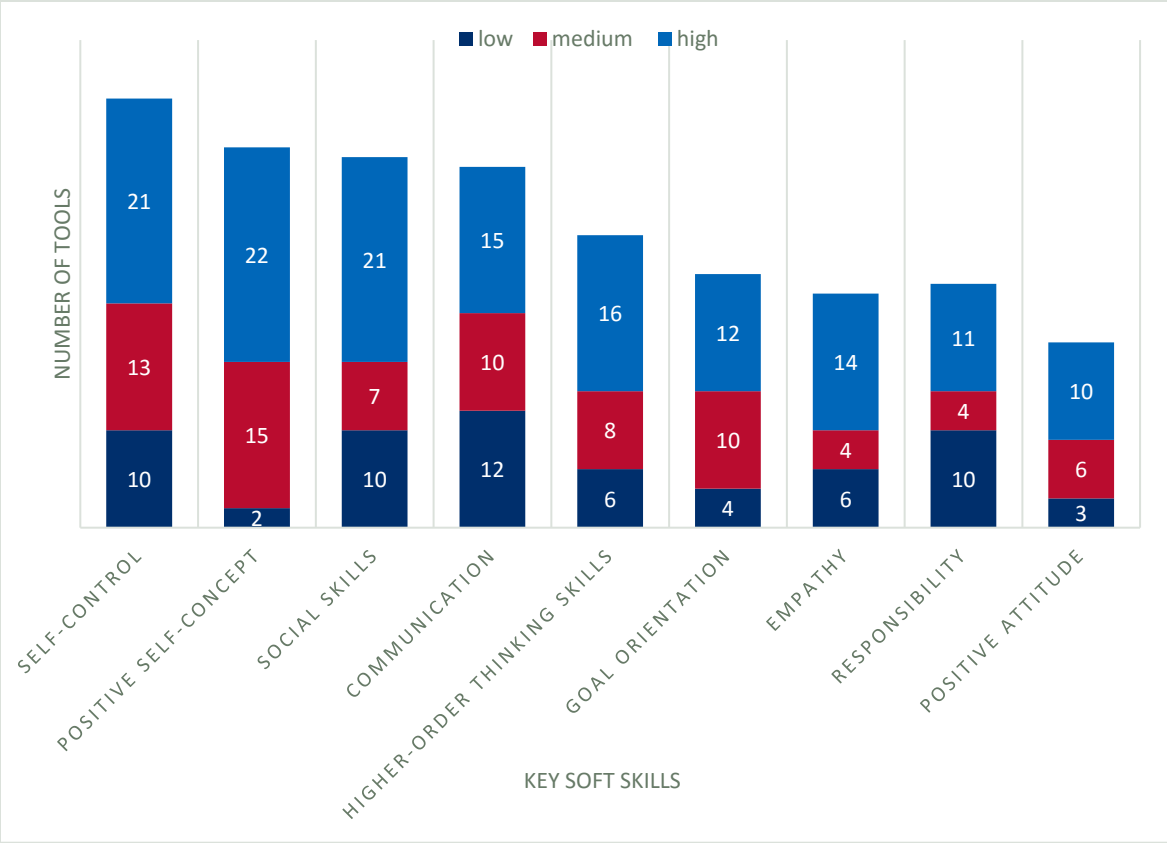
In terms of quantity and quality of measures available, we find that the field of measurement is generally well-aligned with the literature on which skills are most supported by evidence for predicting multiple outcomes. High-scoring tools exist for each of the skills. Overall, the field of measurement remains largely dominated by self-report measures for most of the key skills; the availability of other types of measures (e.g., report by others) is limited and uneven. The challenges resulting from this will be discussed in further analysis below.

[5] The inventory of measures is sortable by skill, for readers who wish to perform more in-depth analyses.

Quantity of Tools Available by Skill

The figure below shows the number of tools available (by score position) for each of the nine cross-cutting skills of interest for our study. Self-control, positive self-concept, social skills, and communication are all most frequently tested by numerous tools in the field. The less frequent assessment of other skills indicates a gap in the tools we currently have available. Below, skills are compared in terms of number of tools available, presented in descending order from left to right.

Figure 3: Number of Tools Measuring Top 9 Cross-Cutting Skills, by Score Position



*The total number of tools in the visual will add up to more than 74 (N) because some tools measure more than one skill.

Overall, the majority of the tools score highly, but differences by skill are apparent. The most commonly tested skill—with the largest number of high-scoring tools—was self-control. Positive self-concept, social skills, and communication were also frequently assessed, each mostly by tools that meet a high number of criteria. Of the key soft skills, positive attitude was tested least often by the 74 tools.

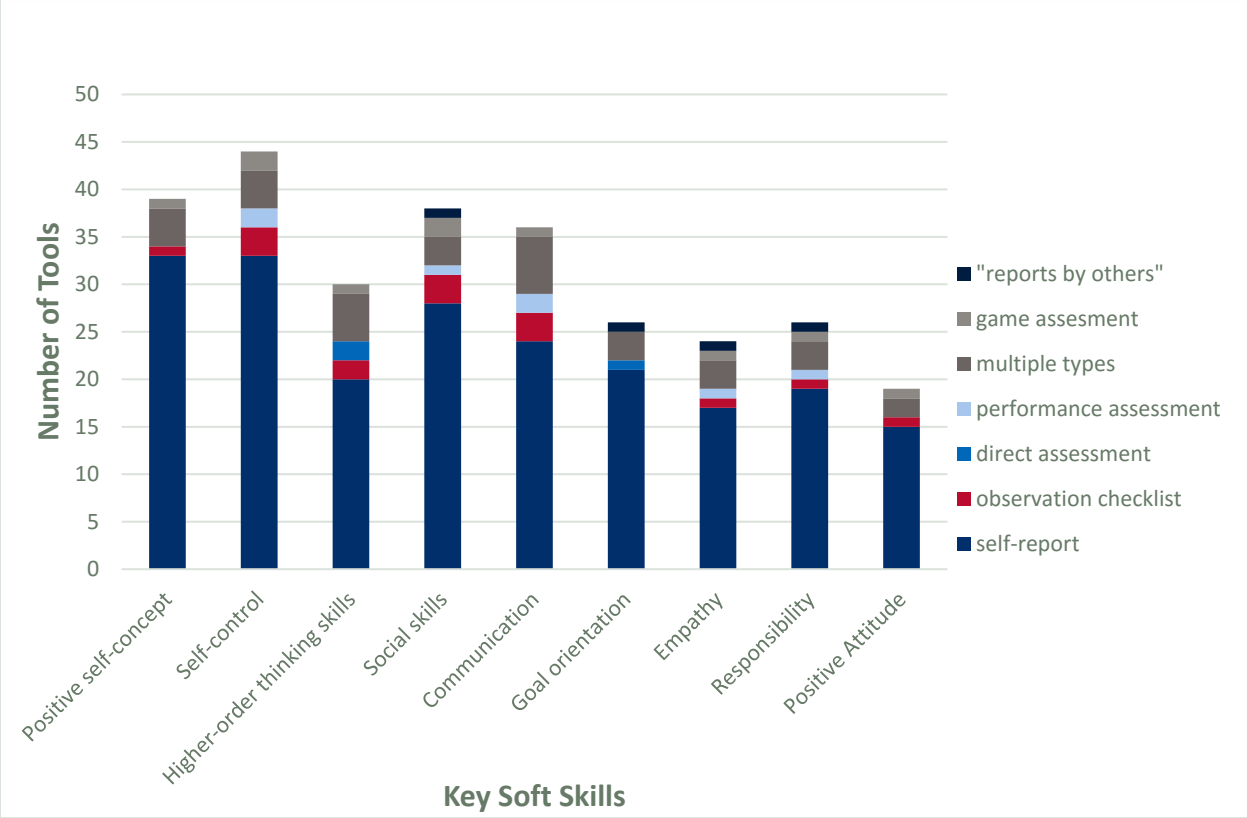
Types of Measures Available, by Skill

Skill measures fall into three main categories: self-reports/self-ratings, reports/ratings by others, and performance assessments and simulations. Under that broad framework, more specific types of tools (as presented in Figure B below) can be categorized as follows:

- Self-reports/self-ratings (including unique methods such as anchoring vignettes, situational judgment tests, and forced choice methods)
- Reports/ratings by others
  - Questionnaires
  - Observational assessments (e.g., observation checklist)
- Performance assessments and simulations
  - Performance assessment
  - Direct assessment
  - Games or simulations

Figure 4 below also contains an additional category for multiple types of assessments, noting instances in which a tool used multiple methods (combining self-report and an observation checklist for example). For further definitions of these terms see the section on the State of the Field beginning on page 12.

Figure 4: Types of Tools for Each Key Skill



Self-reports were overwhelmingly most common across all skill types. Multiple methods of assessment are the next most common, followed by observation checklists, and performance assessments. Some skills (such as social skills and empathy) enjoy a greater diversity of assessment types than others. The dominance of self-report tools, despite their well-known biases and limitations as discussed in the introduction, indicates a gap in the field and a need for measures employing other methods.[6]

Overview of Full Universe of Measures Reviewed

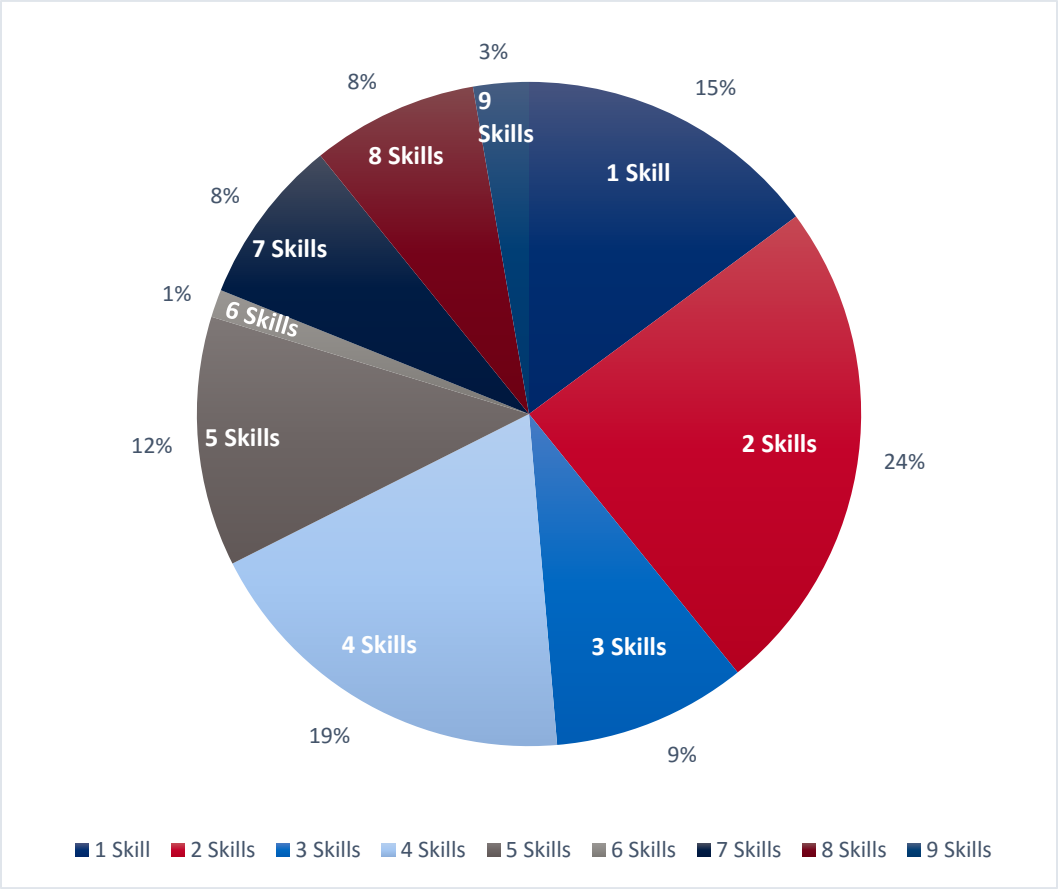
In this section, we present characteristics of the overall universe of tools reviewed. This includes information on how many skills are measured, levels of validity and reliability, outcomes, and larger demographic and geographic information about with whom and where these measures are used.

[6] We recognize that there is good reason for the prevalence of self-report tools; they are quick, inexpensive, easy to use, and potentially reliable (Duckworth and Yeager, 2015, p. 239). Because of their unique vulnerability to measurement error, however, we advocate for their triangulation with additional methods where possible (which we will discuss in more detail in our conclusions and recommendations).

Key Soft Skills Measured

Many programs may wish to use measures that encompass as many as possible of the nine key skills of interest for this review, although it is important to note that more is not always better (as such, number of skills measured was not a scored criterion for this review). For example, a program may only study social skills, and a tool may score well but only focus on social skills. If it were penalized for measuring only one of the skills of interest, the program might miss a tool valuable for that program. We found, however, that only about one-third of tools in the inventory measure five or more skills. The other two-thirds measure between one and four skills. Figure 5 below presents tools according to how many key skills are measured.

Figure 5: Percent of Tools Categorized by Number of Skills Measured



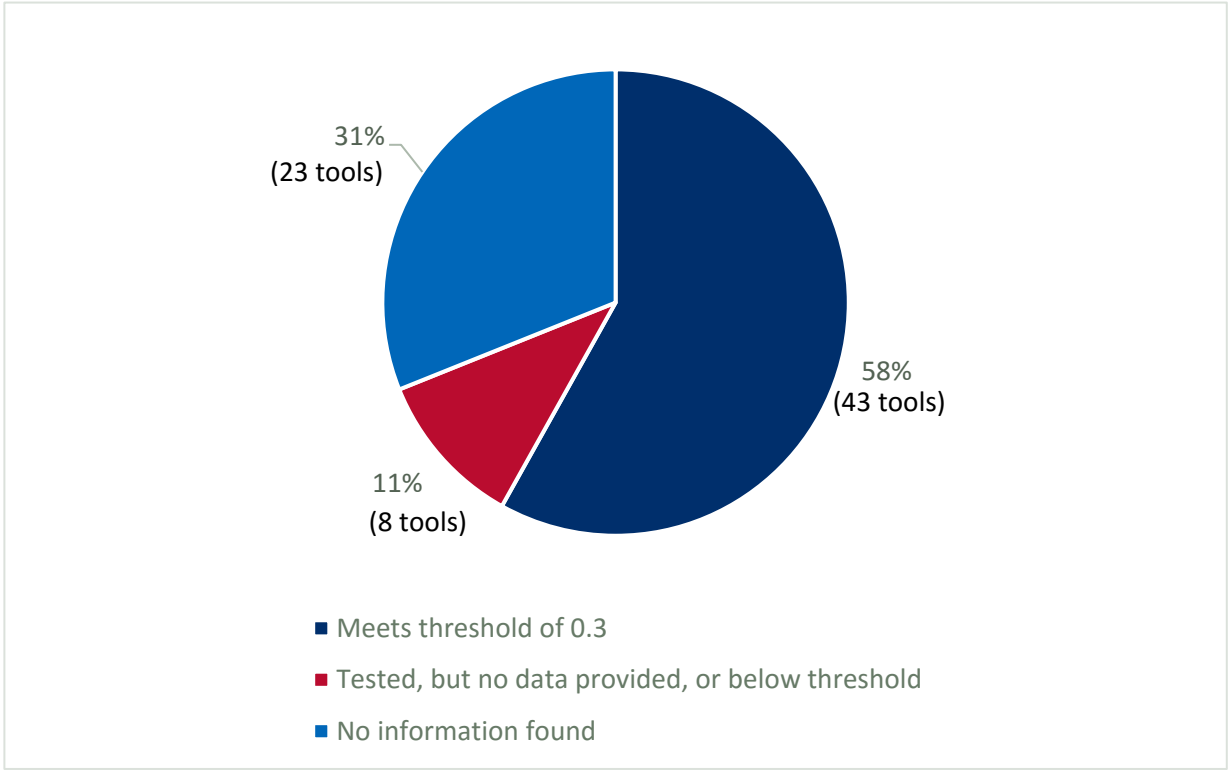
Tool Validity, Reliability, and Outcomes

Measures of a tool's validity and reliability are important—validity demonstrates that a tool measures what it intends to measure, and reliability shows it is doing so consistently across multiple respondents and multiple administrations. Evidence of reliability testing (either meeting an acceptable threshold or evidence of having been tested but without results reported) exist for the majority of tools in the database. This is not the case for validity, however, indicating a need in the field for greater testing and/or documentation of such evidence.

When possible, evidence of validity and reliability is documented in cases where developers or researchers have provided specific test results that meet commonly recognized thresholds, (e.g., values of 0.3 for predictive validity and 0.7 for Cronbach's alpha test of reliability). In other cases, developers report having tested validity and reliability without providing specific test results; such cases are categorized separately, but still receive partial credit for meeting these criteria.

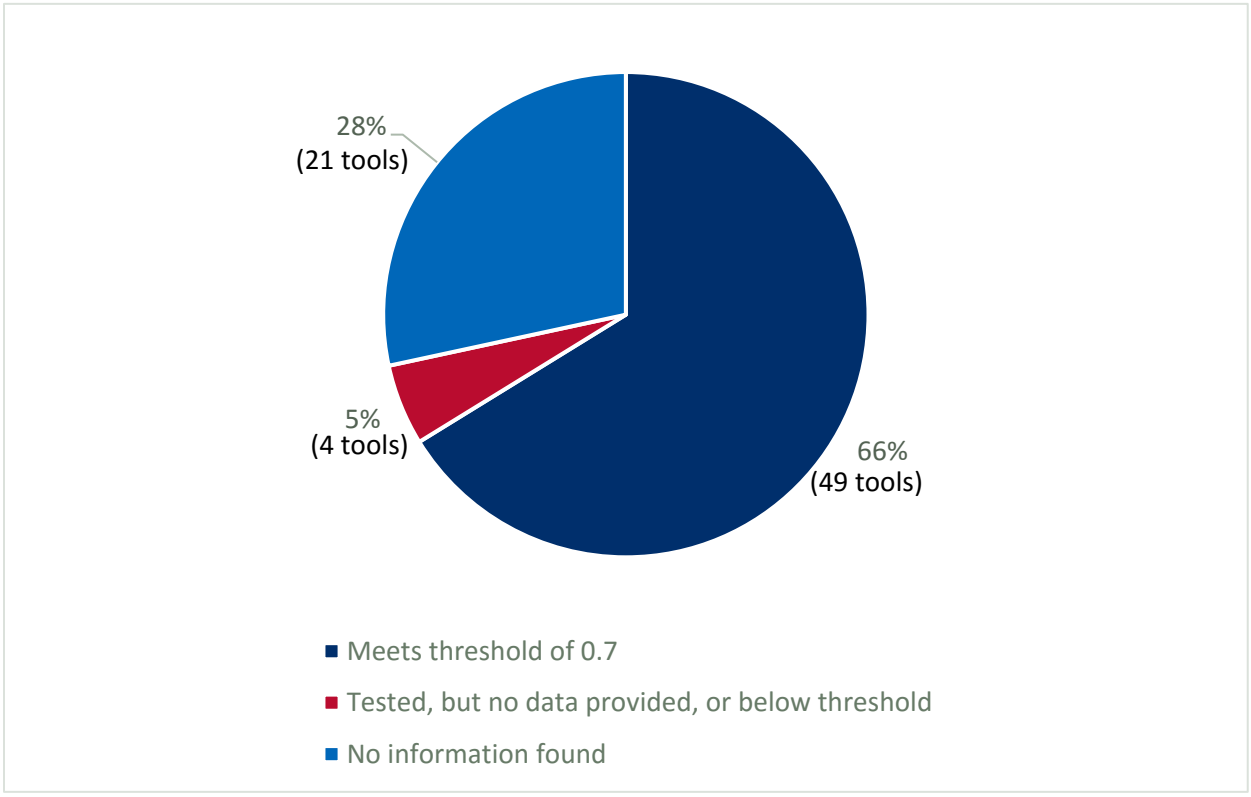
Analysis of the database indicates that 43 tools (58 percent) are proven valid, as demonstrated by coefficients that meet acceptable thresholds for validity from tests using their original validation sample, while 8 tools (11 percent) report validity in their documentation without providing validity coefficients. No information was found for 23 tools (31 percent). The breakdown of tools by evidence of validity is shown in Figure 6.

Figure 6: Percent of Tools by Evidence of Validity



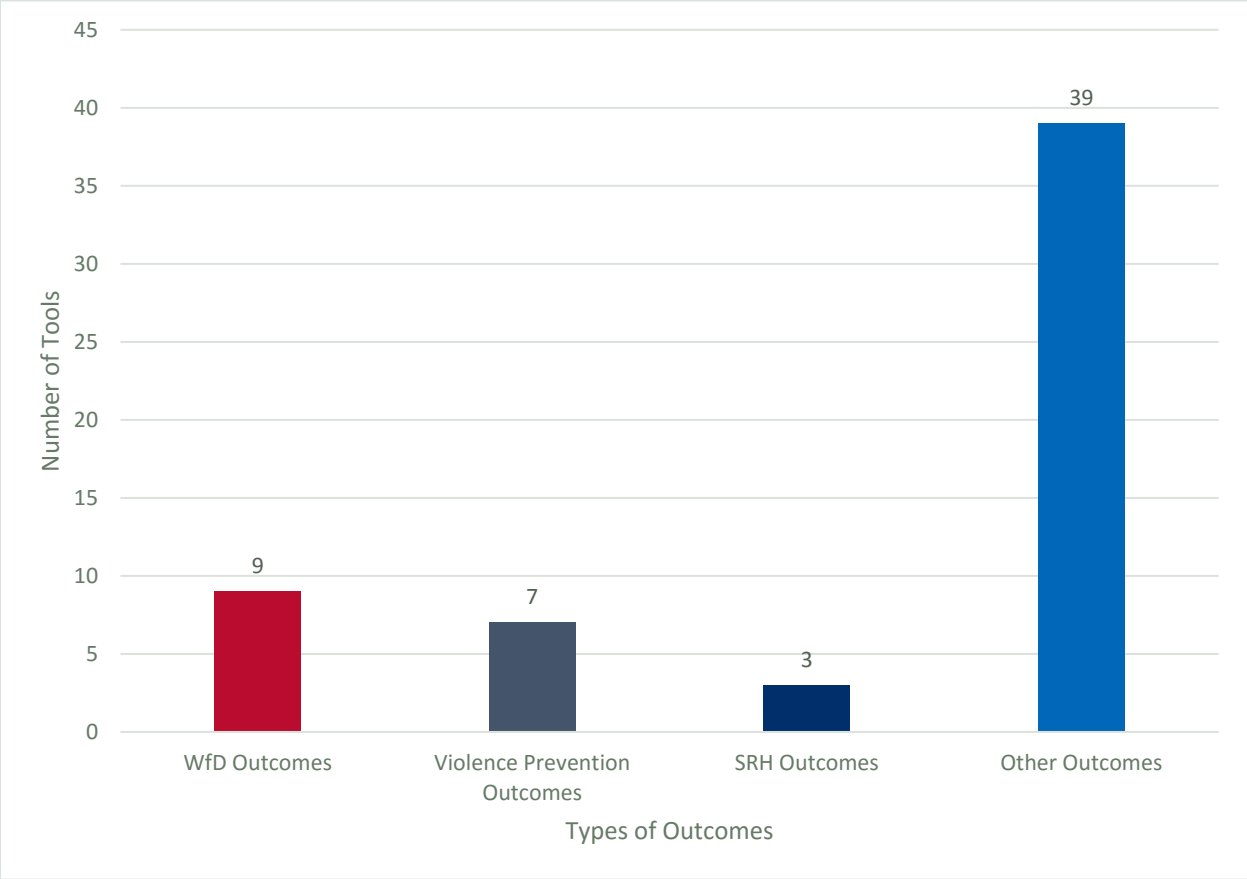
Evidence of reliability was identified for 49 tools (66 percent), which met the accepted threshold of 0.7 for Cronbach's alpha, while developers or researchers reported reliability without providing values for four tools (5 percent). No information was found for 21 tools (28 percent). Figure 7 shows the proportion of tools in each category of tool reliability (shown as a percentage of 74 total tools).

Figure 7: Percent of Tools by Evidence of Reliability



Predictive or concurrent validity of the measures is determined with their relationships to specific outcomes of interest, which is another key indicator of tool relevance. For the purposes of this review, we documented evidence indicating that the tool was correlated with or used to predict outcomes in the domains of workforce development, violence prevention, and sexual and reproductive health. Recognizing that other positive youth development outcomes are also important and relevant, there is an additional category encompassing other outcomes (such as academic performance, prevalence of substance abuse, and more conventional health outcomes). As demonstrated in Figure 8, "other" outcomes make up the largest number of types of outcomes. This is primarily due to the frequent assessment of health (outside of the SRH domain, such as cigarette smoking) and educational outcomes. Figure 8 shows that 41 tools have been correlated with or used to predict an outcome of interest. Because some tools test more than one outcome, however, the total number of cases in the graph is greater than 41.

Figure 8: Number of All Tools that Predict Relevant Outcomes



The review of evidence on validity, reliability, and outcomes reveals distinct gaps in information on key criteria that would enable program implementers to evaluate the quality and relevance of many soft skill measurement tools. Our research team searched for evidence that tools are aligned with USAID or youth programming, meaning that a tool has been used in a particular

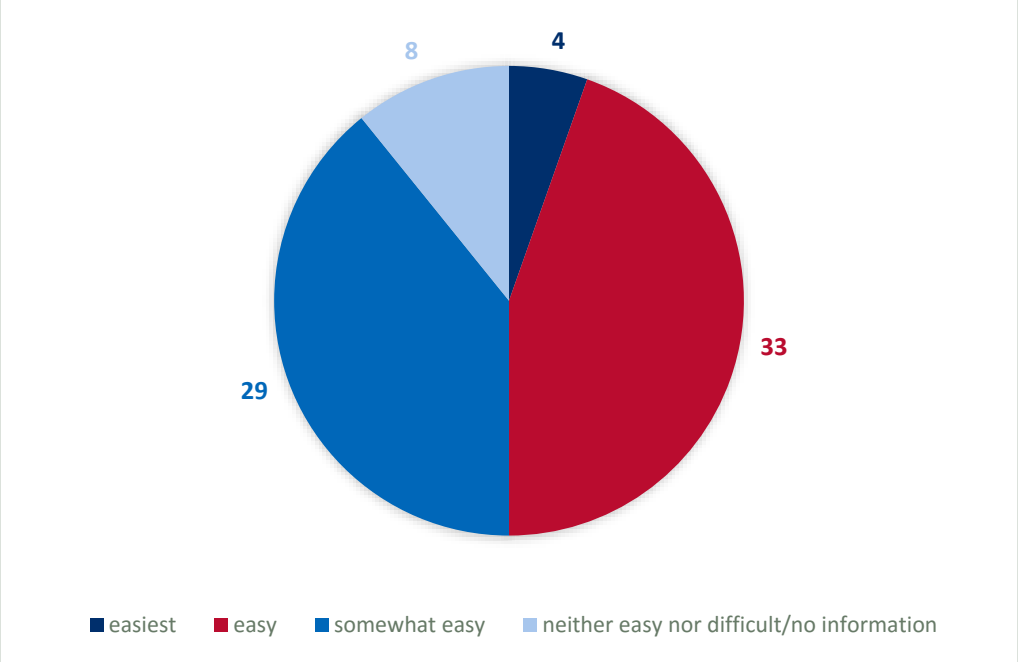
program context and validated with outcomes relevant to youth development programming. Forty-three tools or 58 percent, are aligned with international youth programming purposes, based on these criteria (see Table 4 for criteria definitions). The gap does not necessarily mean such information does not exist, but does show that it is not readily available, even after significant team effort to find it, including direct outreach to tool developers.

Purpose and Ease of Use

During the development of the database, several practitioners noted the importance of how "easy" a tool was to use. In these conversations ease was often described in relation to whether translations of a tool exist, and the extent of preparatory work required in order to use it in a program setting. Another important consideration noted was the degree to which the tool could be administered and answered quickly by participants. (It should be noted here that length of time is considered per construct, not length of time to administer the entire tool.) For scoring purposes, ease of administration was composed of three parts—the need for specifically trained staff (beyond basic knowledge of soft skills) to conduct or analyze the results of the test, the existence of the tool in languages other than English, and the length of time it takes to administer the test.

Elements that appear to make administration and use of a tool easier are awarded fractions of a point, contributing to the overall score of a tool. A tool can have an ease score that falls into one of five categories, based on the level of evidence available for ease of use: "easiest" to use (1), "easy" to use (0.67), "somewhat easy" to use (0.33), and "neither easy nor hard" to use, or "no information" (0). The vast majority of tools fell into the category of "easy" or "somewhat easy" to use. This can be seen in more detail in Figure 9 below.

Figure 9: Number of Tools by Ease of Use Based on Criteria



The review team compared tools' scores on ease of use against scores on validity, and found that the relationship between the two scores was generally positive, suggesting there is not necessarily a trade-off between ease of use and rigor. This is based on the instances of tools with validity scores that meet thresholds, and the presence of information that meets the criteria for ease of use. Another important factor related to ease of administration that programs must weigh when selecting measures is the ease of analysis and reporting of data generated, although this was not part of our scoring criteria.

Change over Time

It is challenging to find information on whether tools are well-suited to measuring change over time in skill levels. The review identified only three tools that had been validated for measuring change over time at the group level (but not at the individual level). It may be the case that tools are valid for measuring change over time at the group or individual level, but they have not been tested and validated for that purpose. In two cases, tool developers explicitly state the tool should not be used to measure change over time at the individual level: the 12-item Grit Scale and the resilience module of the California Healthy Kids Survey.

Demographic and Geographic Use

The age range of potential respondents is another key factor in the relevance of tools to specific development contexts. Design and content of questions or tasks in a tool may be more appropriate for some ages than others, affecting how a participant responds or performs. This directly relates to tool validity and reliability, since such scores are derived from having tested the tool with participants in a particular age group. If a user is outside the appropriate age range, scores for validity and reliably will no longer be relevant.

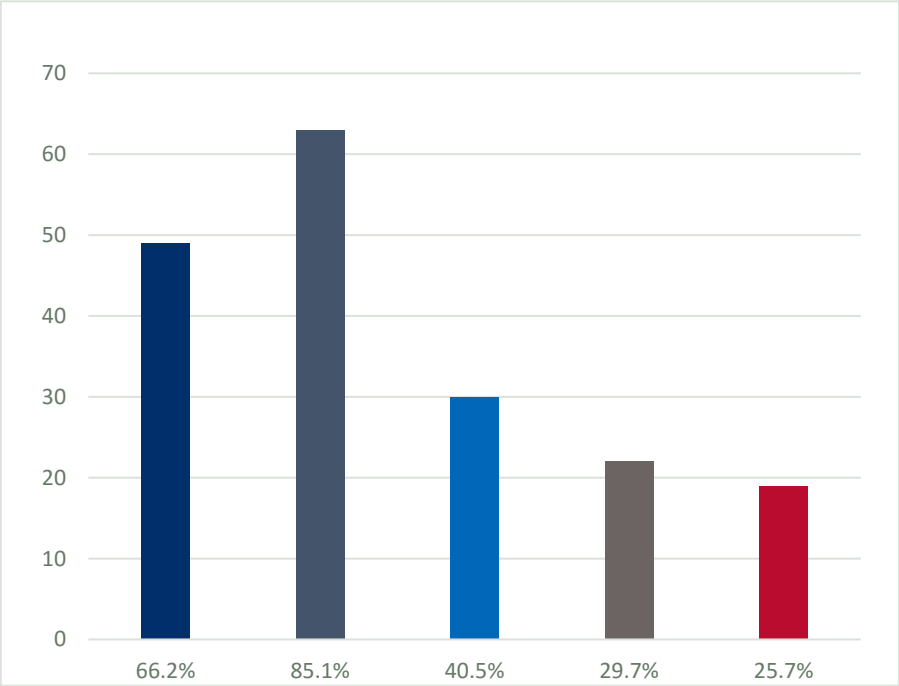
To determine a common set of ages for which a tool might be appropriate, this analysis considered five groups of ages into which a tool's youth population might fall. These are adapted from USAID's Youth in Development policy,[7] and are composed of the following five age ranges: 10–14 years, 15–19 years, 20–24 years, 25–29 years, 30 years and over or adults.

Age cohort information varied widely across the documentation for the tools reviewed. Some tools only reach early adolescents, while others are suitable for almost all ages. Others still did not specify age groups other than through terms such as "young people" or "adolescents," or a particular school or grade range. In those cases, a reasonable approximate age was assigned to the terms, for example "middle school" is commonly used for youth ages 11–13 in higher income countries.

[7] USAID (2012). Youth in Development: Realizing the Demographic Opportunity.

As shown in Figure 10, the majority of the tools have been used to assess soft skills of youth ages 15–19, followed by early adolescents ages 10–14. About 40 percent have been used with youth 20–24, while less than a third have been used to assess young adults in the 25–29 years and age 30 and over ranges, respectively.

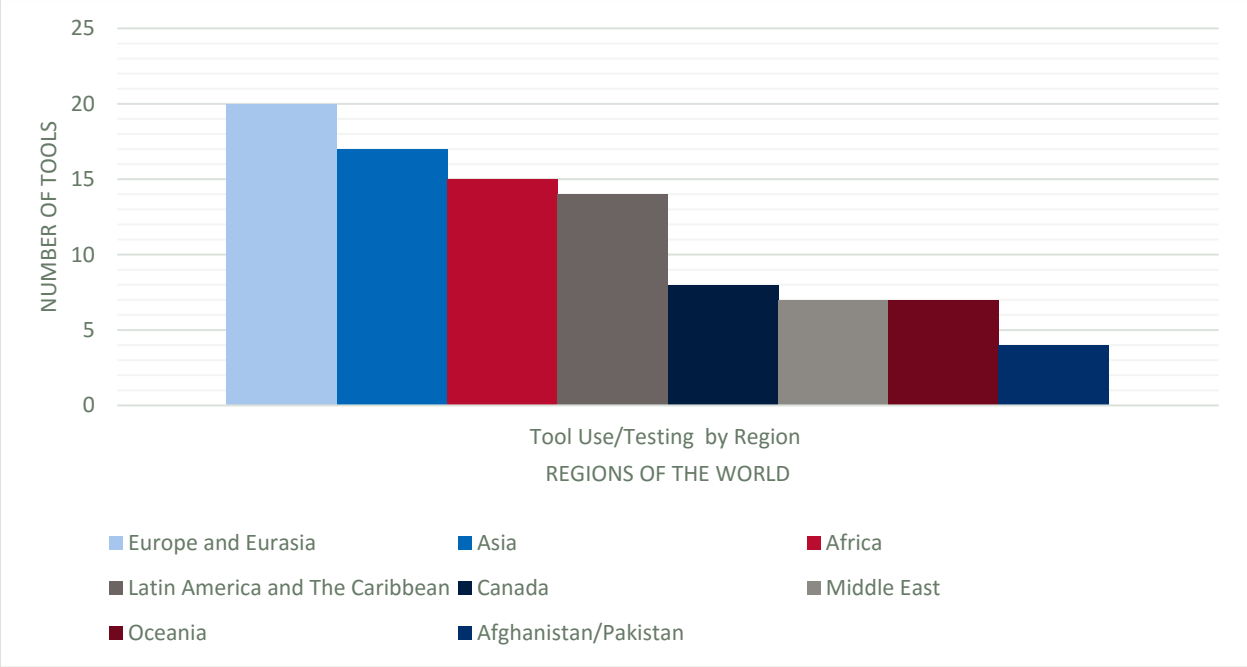
Figure 10: Percent of Tools by Age Ranges Assessed



Locations of Use

Information on where tools have been used before is particularly important for international users, and it has been recorded in the database when available. Developers provide a broad range of location types, however, such as cities, states, countries, or geographic regions of the world. To harmonize such information, this analysis is based on the regional framework used by USAID, plus two additional categories for regions not clearly designated in that framework: Canada and Oceana (Australia and New Zealand). Figure 11 below breaks down tools according to international use in those regions. Overall, 49 out of the 74 tools have been used outside of the United States. A tool may appear in more than one region, based on where it was used. Some tools have been used in more than 10 countries, while others have only been used in one country. The United States was not included in this section since the focus was on youth outside of the United States.

Figure 11: Number of Tools Tested by International Region



Review of Top-scoring Tools that Measure the Key Cross-cutting Skills for Youth Development

The YouthPower Action team conducted a review of 74 instruments to inform the field about a set of tools that addressed the key soft skills emerging from our review of the literature that enjoy strong evidence that they promote positive workforce and sexual and reproductive health outcomes, and prevent violence. Further, we reviewed each tool based upon a set of criteria, described in more detail in the Methodology section of the paper, and then scored each tool according to the degree to which they met the criteria listed below.

- Free and open access
- Appropriate for youth ages 12–29
- Measures skills emerging from our literature review of key skills for cross-sectoral youth development
- Evidence of validity and reliability
- Tested with youth development outcomes relevant to the project
- Tested with relevant validation samples
- Evidence of international usage
- Ease of administration

The tools were then divided into three groups based upon the degree to which they met all of the criteria: high (meeting five to seven criteria), medium (meeting from three to fewer than five criteria), and low (meeting fewer than three criteria).

All of the selected tools from the inventory in Table 5 below score high on the criteria, *and* measure the top three skills found to promote positive cross-sectoral youth outcomes: positive self-concept, self-control, and higher order thinking skills. In addition, they all measure other skills that emerged in our review of the literature as having strong evidence that they foster at least two of the outcome areas: social skills, communication, empathy, goal orientation, plus positive attitude, and responsibility, both of which enjoy a lower level of evidence that they foster all three outcomes. The top scoring tools are listed in Table 5 in order of the number of key skills that they measure, from the top tool measuring nine skills, to the bottom tool measuring four skills.

Ten tools score high on the criteria and measure the top three cross-cutting skills for youth outcomes. As such, these tools may be of particular interest to international youth development programs working to promote positive workforce and SRH outcomes, and preventing violence. Some of the tools have been used in conjunction with other outcome areas as well, such as education, psychological and emotional health, substance abuse, and health (see the inventory for more details). The strengths and weaknesses of each tool are discussed below.

Some general observations can be made on this set of tools. The first is that the tools aligned in their measurement of the nine skills. As in the research literature reviewed in the report, “Key Soft Skills for Cross-sectoral Youth Outcomes,” there is some divergence in the skills measured by these instruments after the three most commonly measured key skills that foster cross-sectoral outcomes (positive self-concept, self-control, higher order thinking skills). Social skills are the next most commonly measured skill, measured by all tools except one (the Responses to Stress Questionnaire). Communication, empathy, and positive attitude are assessed by seven tools, followed by responsibility, and lastly, goal orientation.

TABLE 5: TOP SCORING TOOLS THAT MEASURE KEY SKILLS OF INTEREST

Assessment Name	Tool Score Position	# of Key Skills Measured	Positive self-concept	Self-control	Higher-order thinking skills	Social skills	Communication	Goal orientation	Empathy	Responsibility	Positive attitude
California Healthy Kids Survey, Social and Emotional Health Module	high	9	√	√	√	√	√	√	√	√	√
Chinese Positive Youth Development Scale (CPYDS)	high	8	√	√	√	√	√	√	√	√	
SENNA 1.0	high	8	√	√	√	√	√		√	√	√
SENNA 2.0	high	8	√	√	√	√	√		√	√	√
Child and Adolescent Wellness Scale (CAWS)	high	7	√	√	√	√		√	√		√
The Anchored BFI Tool	high	7	√	√	√	√	√		√		√
The Big Five Inventory	high	7	√	√	√	√	√		√		√
Knack	high	6	√	√	√	√				√	√
Jamaica Youth Survey	high	5	√	√	√	√		√			
Responses to Stress Questionnaire (RSQ)	high	4	√	√	√		√				

The second observation relates to the purpose of the tools. Evidence showed that just two tools have been explicitly used for youth program evaluations: the Chinese Youth Development Scale and the Jamaica Youth Survey. The Chinese Youth Development Scale was used with the P.A.T.H.S. project in Hong Kong to measure holistic positive youth development, while the Jamaica Youth Survey was used in an evaluation of violence prevention programs in Jamaica. Another group of tools has been used in school systems: The California Healthy Kids Survey, Social and Emotional Health Module, and the SENNA 1.0 and 2.0 instruments have been used to monitor social and emotional skills at the aggregate level across large populations of students; although they are very effective for that descriptive monitoring purpose, they are not recommended for use in evaluating the effect of programs on individual participants' performance, or for "high stakes evaluations" that result in consequences for individuals or programs or schools (for example, promotion, salary increases, program funding). It is unknown whether those survey items would be sensitive enough to measure change over international youth development programs of short duration. The third group of tools are used for psychological assessments, including the Child and Adolescent Wellness Scale, the Big Five Inventory (BFI) and the Anchored BFI, and the Responses to Stress Questionnaire. They have been used with a variety of outcomes relevant for positive youth development programs and enjoy the highest level of validity and reliability, but again have not been validated for measuring change over time among individual youth program participants. Finally, the Knack game is an application that requires a computer or smartphone with embedded assessments of soft skills that is used by individuals, and conceivably, could be used in conjunction with a coach or mentor to help youth develop their skills. It is unknown whether the app would detect changes in performance that could be tied to a program intervention.

The third observation is that all of these instruments use self-report by youth, with the exception of the Knack game, which is a performance assessment. The Chinese PYD Scale documentation also suggests that it could be used for reports by others as well as youth, which can be helpful to triangulate responses to reduce bias in reporting. As noted above, self-reports are subject to both social desirability bias or "faking" answers that are perceived as desirable by the youth, as well as reference bias, or the tendency to compare oneself to those in one's immediate social reference group, and thus, may not be objective. There is general recognition of a tendency toward an upward bias in reporting on all measures of positive attributes or behaviors, and this is also problematic in trying to capture positive change over time during a program since there is little room for improvement if self-reports start out high (Lippman et al., 2014). Also, youth may score themselves lower at the end of a program after gaining perspective and understanding of the skills, when in fact the program has actually succeeded in improving their skills. Since the highest scoring tools addressing key skills use self-report as the method of assessment, as do the majority of tools in the inventory, it suggests the need for improvement in the field in order to objectively measure skills in youth over time.

In-Depth Review of Top Scoring Measures

We review each measure (or group of measures, when closely related), in the order presented in Table 6 below. For each measure, the table shows the results according to each scoring criteria as well as the overall score. The full set of measurement tools and their associated characteristics as well as detailed information on the tools described can be found in the inventory.[8]

[8] Available for download from www.youthpower.org as a companion to this report.

TABLE 6: HIGH SCORING TOOLS: SCORING CRITERIA RESULTS

Assessment Name	Evidence of use in Int'l. Youth Programs	Evidence of Validity	Outcomes Tested	Evidence of Reliability	Evidence of Int'l. use	Relevant Sample Tested	No Trained Personnel Required	English + Other Languages	Short Length of Admin.	TOTAL (out of 7.0)
California Healthy Kids Survey: Social and Emotional Health Module	yes	yes*	yes	yes*	yes	yes	not required	yes	no info	6.67
Chinese Positive Youth Development Scale (CPYDS)	yes	yes*	yes	yes*	yes	yes	not required	yes	no info	6.67
SENNA 1.0	yes	yes*	no info.	yes*	yes	yes	not required	yes	no info	6.67
SENNA 2.0	yes	yes*	no info.	yes*	yes	yes	not required	yes	no info	6.67
Child and Adolescent Wellness Scale	no	yes*	yes	yes*	yes	yes	not required	yes	no info	5.67
The Anchored BFI Tool	yes	yes*	yes*	yes*	yes	yes	not required	yes	no	6.67
The Big Five Inventory	yes	yes*	yes*	yes*	yes	yes	not required	yes	yes	7
Knack	yes	yes	yes*	yes	yes	yes	not required	yes	no	6.17
Jamaica Youth Survey	yes	yes	yes*	yes*	yes	yes	not required	yes	no info	6.67
Responses to Stress Questionnaire (RSQ)	yes	yes*	yes	yes*	yes	yes	no info.	yes	no info	5.33

California Healthy Kids Survey (CHKS), Social and Emotional Health Module (SEHM)

The California Healthy Kids Survey, SEHM incorporates measures of all nine key soft skills in a short 21-item format. The module is part of the California Healthy Schools Survey, used by school districts and schools in California to measure, through student self-report, the strengths of students ages 10–19. It has been used both for monitoring and evaluation. For the purpose of monitoring, it has been used to track district- and school-level trends in students' well-being. Survey results have helped to better plan and target social and emotional learning interventions. Another highly relevant element of the survey is the Resilience and Youth Development module, which has more complete short scales for goal orientation, communication, and problem solving. The school climate module has a short scale for social skills that is more elaborated than the one in the SEHM.

The survey enjoys strong psychometrics in terms of reliability and validity (see the inventory for details), and has been tested in large and diverse school districts in California, with evidence that it has tested well among both males and females, and among blacks, Latinos, Asians, and whites. It has shown to be predictive of school and quality of life outcomes. Like most measures reviewed for this inventory, it has not been recommended for use in measuring change over time among individual students. International versions are available in Spanish, Japanese, and Korean, and the short, simple items would be easy to translate to additional languages. It received a high score rating of 6.67. The response options are a 4-point Likert scale, from "not at all true of me" to "very much true of me." The limited range of the scale provides limited ability to discriminate at the high end and may provide results with an upward bias and limited room for improvement over time. This survey is easy to administer with paper and pencil.

The SEHM is useful for programs seeking to assess youth quickly at one point in time, to get a sense of their general self-perception on these nine skills in comparison with others in a program. Data could be aggregated to the program level and could also be used to compare across programs.

Chinese Positive Youth Development Scale (CPYDS)

CPYDS has been used to assess the effectiveness of a large-scale positive youth development program in Hong Kong, Project P.A.T.H.S. It is based upon positive youth development and positive psychology theory, and has been used in Hong Kong and Macau among youth ages 12–18. It addresses eight of the nine priority skills for cross-sectoral youth development, including: positive self-concept, self-control, higher order thinking skills, social skills, communication, goal orientation, empathy, and responsibility, and many others as well. It also scores very highly—6.67—on our criteria, including evidence of reliability and validity, and it has been tested with outcomes of interest to cross-sectoral youth development programs; it is negatively related to delinquency, problem behavioral intent, and substance use, and positively related to thriving, wellness, and life satisfaction. It is a self-report questionnaire with 90 items and yes/no answers. The developer acknowledges that it could be improved with the addition of reports by others, as well as further examination of subscales. As one of the most

comprehensive assessments reviewed, as well as one of the highest scoring, and with its stated purpose and prior use in assessing youth development programs, this instrument is highly appropriate for use by international youth development programs.

SENNA 1.0 and 2.0 Surveys

The SENNA instrument (Social and Emotional or Non-Cognitive National Assessment) was developed with funding from the Ayrton Senna Institute to assess social and emotional skills among school-aged children and youth in Brazil, for the purpose of education system-wide monitoring and evaluation. It has been used for 5th, 10th and 12th graders in Rio de Janeiro and other municipalities in Brazil, and has been tested with over 24,000 children and youth. The instrument combines measures of the Big Five Personality Factors: Conscientiousness, Extraversion, Agreeableness, Emotional Stability, and Openness to New Experiences. It also includes a measure of Locus of Control, which was found through factor analyses to be a necessary addition to their conceptual framework, which was developed from analyses of existing scales selected according to criteria for adaptation for the study. Measures can be found in the instrument of eight of the nine priority skills emerging from the YouthPower Action literature review, "Key Soft Skills for Cross-Sectoral Youth Outcomes." In addition to the top three, those include social skills, communication, empathy, responsibility, and positive attitude.

The instrument is easy to administer, with students self-reporting on each item using a 5-point Likert scale from "not at all" to "totally" for 76 items, along with three anchoring vignettes (see explanation of these below under the Anchored BFI) that improve the instrument's reliability and validity. As a result, the instrument enjoys high levels of validity and reliability, and scores on the instrument have been found to be related to grades in mathematics and Portuguese. It is available in English or Portuguese for free with permission from the Senna Institute. This instrument is appropriate for monitoring the performance of classrooms or schools or youth development programs, but authors state that it is not appropriate for individual assessment or program evaluation. Group-level scores such as those produced by this instrument can be used to monitor educational systems or programs, without making conclusions on individual performance in specific skills, or whether a program is successful in changing those skills at the individual level. The cautions by the authors are intended to prevent an invalid use of the instrument, to prevent conclusions to be drawn that are not supported by the instrument's design. It received a high score of 6.67 points according to our criteria. Limitations of the instrument noted by the authors include the need to further specify skills related to facets within the broader Big Five-based framework, with measures that are sensitive to change over time.

SENNA 2.0, released in December 2016, is an improved version of the SENNA survey (specifically, the authors added items to measure self-efficacy, making the tool more complete) that is appropriate for use for summative and descriptive purposes of social and emotional skills in a group population. The authors state, however, that it is not appropriate for "high stakes" testing that has consequences for individual promotion, teacher incentives, or program funding, for example. It is a self-report instrument with Likert-type scales. A short version is available for youth ages 12–14, as well as a longer version for those ages 15–19. It has been successfully tested on a much broader population of students in school in Brazil, and results have been

analyzed in relationship to academic achievement outcomes. It is available in Portuguese and English. The instrument measures 17 facets of social and emotional skills organized into five broad domains that correspond to the Big Five Personality Factors but with different names; this framework evolved based upon testing among Brazilian students: Openness to the New, Amity, Self-Management, Emotional Resilience, and Engaging with Others. It includes measures of behaviors in daily life in these areas, as well as measures of how well the student thinks that he/she performs on these domains, that is, measures of self-efficacy in each domain. Measures of the same eight key soft skills that were found in SENNA 1.0 that emerged from our review of the literature are found in SENNA 2.0.

The developers expect that the measures of self-efficacy in each domain would be the first to respond to change in skills as a result of program interventions, followed by behavioral measures. Information on validity and reliability was informally considered for this review, since results are in the process of being published.

Child and Adolescent Wellness Scale (CAWS)

The Child and Adolescent Wellness Scale (CAWS) is a self-report questionnaire that was designed to assess strengths, specifically, adaptive qualities, in school-aged children that indicate psychological health. It is based upon theory from positive psychology, resilience research, and prevention science. It has been validated for use among adolescents ages 12–18, with samples in the United States, Japan, South Korea, Taiwan, and Thailand. It includes seven of the nine skills that are being targeted by this review, including self-control, positive self-concept, higher order thinking skills, social skills, empathy, goal orientation, and positive attitude. Students respond to 150 items with Likert-type response scales from “strongly disagree” to “strongly agree.” It scores 5.67 on criteria for this study, including reliability and validity, as well as ease of administration. It can be used for assessing individual youth to help them identify areas of strengths and areas that need further development for their psychological health.

The Big Five Inventory (BFI) and the Anchored BFI Tool

The Big Five Inventory, or BFI, is one of the most-used instruments worldwide for identifying theory-based individual personality factors, and for comparing the consistency of those factors across cultures. The Big Five factors are: openness to experience, or the capacity to enjoy “new” ways of thinking about the world; conscientiousness, or the propensity to organize and achieve; extraversion, or the propensity toward social interaction; agreeableness, or positivity in interactions with others; and neuroticism (or its opposite, emotional stability), the ability to manage stressful situations or emotions (John, Donahue and Kentle, 1991). Each of the Big Five factors has six facets within each factor. Although there is variation in the field on how to define each factor and its facets and how each facet corresponds to skills, an emerging literature has used factor analysis to examine various measures and help define more universally the facets and corresponding skills subsumed under each factor. Lippman et al. (2015) provided a crosswalk from all of the Big Five factors to soft skills based upon this literature, using the common skill terminology developed for that report and that continues to be

used for this project. For example, the Big Five factor of Conscientiousness includes facets that relate to the skills of being organized, hardworking and dependable, self-motivated, having self-control, and integrity/ethics. Seven of the priority skills for cross-sectoral youth outcomes that are assessed by the BFI include the top three: positive self-concept, self-control, and higher order thinking, plus social skills, communication, empathy, and positive attitude, albeit some are very thinly assessed with only one or two items.

One critique of using personality assessments for monitoring skill development is that personality factors are considered enduring individual tendencies that may not be malleable to program interventions. Current research on personality attributes such as the Big Five shows that they themselves are malleable over the life course, so that the idea of personality being an unmalleable "trait" no longer has currency, and the skills that are expressions of the facets are malleable through interventions (Roberts et al., 2006)

The BFI is a 44-item simple assessment that employs a quick response layout and takes only five minutes. It has been administered to those ages 10 and over in many countries around the world, and it has been translated into 28 languages. It uses a Likert scale with five response options from "strongly agree" to "strongly disagree." It is one of the few instruments reviewed that meets all four criteria for ease of use. It enjoys high levels of validity and reliability, and has been found to predict outcomes in all areas of interest, including workforce, violence prevention, sexual and reproductive health, and education, health, among others. It earns 7 points in meeting our criteria. The limitation for use by programs is the need for more research on the sensitivity of its items to change over the duration of typical youth development programs.

The Anchored BFI uses the same BFI instrument as above, but with the addition of anchoring vignettes, or scenarios, as well as situational judgment tests, presented to the respondent, to address the issue in which respondents in survey research have different underlying response patterns, or different standards or cultural biases, and those patterns are a hidden part of the variation between people taking a survey (Pagel et al., 2016). By asking respondents to rate hypothetical individuals' soft skills, those hidden patterns become explicit, and then the BFI responses can be adjusted accordingly. This then improves the survey's ability to correct for incomparability between individuals from different cultures and countries. Anchoring improves the BFI's reliability, which is useful for international comparisons. It has not been validated for use yet to measure change over time in the Big Five Personality Factors. It earns a high score of 6.67, meeting most of our criteria. The anchoring vignettes do, however, require a higher level of literacy of the respondent, and more sophisticated administration and analysis.

Knack

Knack is a game application that has embedded assessments of soft skills, including the top three skills emerging in our literature review, positive self-concept, self-control, and higher order thinking skills, plus social skills, positive attitude, and responsibility, as well as others that are relevant to work readiness. It is intended to provide youth ages 17 and over with a personal assessment of their skills, and a connection to job opportunities. It has been translated into multiple languages and has been used throughout the world. It receives a score of 6.5 on our criteria, including validity and reliability, and has been tested with workforce as well as

educational outcomes. There are three games, and each game takes 10 minutes, for a total of 30 minutes. It is free, but it does require access to devices with an IOS or Android operating system. It is appropriate for individual use, or perhaps use with a coach or mentor to discuss and improve performance but it is not designed to provide aggregated group scores that would be useful for program implementation or evaluation purposes.

The Jamaica Youth Survey

The Jamaica Youth Survey was designed for use as part of an evaluation to assess the individual level impact of youth development programs in urban Jamaica. It is a self-report survey used for youth ages 12–18, with standard Likert-style response scales, and it addresses five out of nine of the key skills for cross-sectoral youth outcomes, including the top three: positive self-concept, self-control, higher order thinking skills, as well as social skills and goal orientation. It enjoys good reliability and validity, and has been used to evaluate the effectiveness of violence prevention programs in Jamaica, measuring violence and aggressive behaviors. It receives a high score of 6.75 in terms of meeting our criteria. It has the same limitations as other self-report surveys, including social desirability and reference biases by respondents.

Responses to Stress Questionnaire (RSQ)

The RSQ is used to measure coping and involuntary stress responses to a range of stressful situations. It addresses the top three skills for cross-sectoral youth outcomes, plus communication, for a total of four skills of interest. It is a self-report questionnaire that has been used for youth ages 9 and over in international contexts. It can also be used by parents or program staff to report on a youth's behavior. It enjoys excellent validity and reliability and has been used in conjunction with emotional, behavioral, and health outcomes. It receives a score of 5.25 in meeting our criteria. It may be especially helpful to programs particularly in conflict zones or violent areas that are trying to improve youth's response to stressful situations.

7. Conclusion and Recommendations

Youth development program funders, practitioners, and other stakeholders have expressed an urgent need for measures that can reliably assess key soft skills at an individual level over time, within a program implementation context. Our review of existing measures has focused on instruments available to measure a core set of soft skills that have been shown to foster positive outcomes across the domains of workforce development, violence prevention, and SRH, as well as other domains important to positive youth development, such as education, psychological and emotional health, substance abuse, and health (see the inventory for more details). In particular, we have focused on three soft skills with the strongest evidence for cross-sectoral importance: higher order thinking skills, self-control, and a positive self-concept.

Programs will need to evaluate the tools in this inventory for their own purposes. Considerations may include whether the skills being addressed by the program match the skills that are measured in the assessment under consideration, whether the tool has been validated for use with youth of the same age, whether it enjoys acceptable levels of validity and reliability, whether the tool has been used for the same purpose as is envisioned by the program and has been used to measure an impact on outcomes of interest to the program. Other criteria used in the inventory were chosen because of their relevance for assessing skills in international youth development programs and the needs of implementers and evaluators, such as evidence of international use and ease of administration, but not all criteria will be of equal importance to every program.

An ideal tool for evaluating a program's effectiveness in improving soft skills among its participants would:

- Measure all of the three key skills identified as most important across sectors
- Be suitable for measuring change over time in skill levels for either groups or individuals within a program context, depending on the evaluation design
- Meet other criteria of key importance to international youth programs (including ease of administration, validity and reliability, evidence of correlations with positive youth development outcomes, age-appropriateness, free and open access, and evidence of international usage)

We identified 74 existing tools out of a pool of 244 that measured at least one of nine soft skills that have strong evidence that they are linked to at least two of the outcome areas of interest (see "Key Soft Skills for Cross-Sectoral Youth Outcomes"). We reviewed each instrument and scored it according to the degree to which it met each criterion (high, medium, and low), and noted which skills it measured by coding the instrument according to a common skill terminology developed for "Key 'Soft Skills' that Foster Workforce Success: Toward a Consensus Across Fields." This set of 74 tools is included in the inventory.

We then extracted 10 tools that scored highly against our criteria, and measured the top three skills that are linked to all three outcomes areas: higher order thinking skills, positive self-concept, and self-control. In two cases, this group includes different versions of the same tool

(SENNA 1.0 and 2.0; the Big Five Inventory and the Anchored BFI). We described these 10 tools in depth from the perspective of their utility for international youth development programs. They can be grouped in three general categories, as follows.

Program evaluation: The Chinese Positive Youth Development Scale and the Jamaica Youth Survey meet key scoring criteria and have been used to evaluate international youth development programs. The Chinese PYD Scale has the advantage of assessing eight of the top nine skills, whereas the Jamaica Youth Survey assesses five.

Group performance monitoring: The California Healthy Kids Survey, Social and Emotional Health Module, and the Brazilian SENNA surveys are instruments of excellent quality that are useful for monitoring group performance for summative, descriptive purposes, and which have been used in schools and school districts. They are not, however, intended to be used for evaluations that seek to measure individual changes in soft skills over a program's duration.

Individual assessments: The rest of the tools can be used for individual psychological or skill assessments and have been shown to be correlated with outcomes of interest. They can be used in formative assessments in which program staff give feedback and coaching to youth participating in the programs, and may be informative for improving the targeting of skills within a program and for program implementation purposes. They are useful for detecting differences among individuals in a program at one point in time, but they may not be sensitive enough or validated for detecting changes in individuals over the duration of a youth development program.

The field of tools reviewed, even those 10 noted above, suffers from weaknesses and limitations, described below, that obstruct their usefulness for program monitoring and evaluation. In addition, some challenges affect the ability to build evidence in the field *across* programs, which is essential in order to learn what is working and which programs need to be scaled up. Several challenges need to be addressed by the field, as described below.

Terminology: The lack of a common terminology and skill definitions across measurement instruments hampers the ability of program implementers and evaluators to choose instruments that match the set of skills addressed by programs, and to compare results across programs. It also hampers the ability to build the evidence across countries, cultures, research disciplines, policymakers, funders, and practitioners. A proposed common terminology and skill definitions that would bring coherence to the field were suggested in "Key 'Soft Skills' that Foster Youth Workforce Success: Toward a Consensus Across Fields," which was drawn from the research terminology across fields and studies, but also with attention to the terms used by youth, practitioners, and employers.

Evidence of reliability and validity: As noted in the analysis, many tools lacked evidence of reliability and validity, which are essential to provide confidence in the tools. Developers need to be encouraged to publish the results of their tests with their validation samples, and those who have used the tool for assessing youth along with outcomes need to be encouraged to report their reliability and validity.

Self-report: All of the top 10 tools—except the Knack game—and most of the tools in the inventory use youth self-rating scales, which suffer from reference and social desirability biases. It is known that there is a tendency in most cultures to rate oneself at the high end of a scale on a socially desirable quality, as well as to rate oneself high in reference to one's own group. These tendencies not only bias results, but obstruct accurate comparisons across participants in a program, or across programs and cultures, and across time. Using reports by others along with self-reports, and focusing items on actual observable behaviors rather than endorsements of statements, can produce more objective results (Blades et al., 2012; Center for the Economics of Human Development, 2015). For example, the Flourishing Children Project's Goal Orientation scale includes the following behavioral item, "How often do you make plans to achieve your goals?" on a frequency scale from "none of the time" to "all of the time."

In addition, anchoring vignettes and situational judgment tests have been successful in reducing these biases and increasing validity and reliability, but require a more sophisticated and costly administration and analysis process, and a high level of literacy of respondents. Further, as discussed in the introductory section on page 13, anchoring vignettes have not yet been validated to detect change over time.

Response scales: Response scales are often overlooked in reviews of instruments, but they are critical in determining the sensitivity of items to detect differences between program participants and within participants over time. Most of the instruments reviewed use simple Likert scales, which are good for identifying differences in general tendencies between individuals, but finer grain response scales are needed. Specifically, improved response scales could address the tendency toward an upward bias in self-report, by capturing variation at the upper end of scales to differentiate between youth who really excel at a skill and those who are just above average (Lippman et al., 2014). Making such distinctions could establish thresholds that could help answer the question of how much of a skill is enough to affect an outcome. Finer grained responses at the upper end also allow for the detection of growth over time within an individual, due to a "ceiling" effect. If a youth rates highly at the start of a program, there is no room on the scale to detect growth. Measuring frequencies of behaviors, when possible, is more objective than endorsement of the skill by the youth, and can be used in "other" reports as well (Lippman et al., 2014). When youth reports are triangulated with measures by others for more objectivity, it raises the additional challenge of making sure that both youth and adults or "other" reporters share the same concept/understanding of the skill, which is, of course, essential to model and develop the skill among youth.

Developmental appropriateness: There are differences in how skills manifest as youth age. The age span from 12–29 is large and encompasses huge differences in development, including cognitive processing, identity formation, emotional regulation and executive function, social contexts, life experiences, and academic, technical, physical, and practical skills, to name a few. Items need to be used, adapted, or developed that are appropriate for specific age groups and that reflect the youth's understanding of a skill and how it is demonstrated across contexts and relationships, such as school, work, with peers, or with family members. Most measures found were for adolescents ages 15–19 rather than early adolescents or young adults, and so will need to be adapted or developed to suit all age groups of interest.

Measuring change over time: Research is needed on how to reliably measure change in soft skills over time. This is both a key area for future investment and a difficult task; it is needed for program evaluations that seek to determine whether a program has been successful in improving skills among youth, but few have been validated for that use. Some assessment developers warn against using their measure for high stakes accountability, as noted in our analysis. Programs can succeed in educating youth and raising awareness about what is involved in a skill, and giving youth practice in using a skill, yet scores can decline in the program as a result. This can happen as youth develop a more accurate self-perception of their skills in relation to others and to their own potential. The use of frequencies of behaviors and "other" reports as well as anchoring vignettes may help to more accurately measure improvement.

Validation of instruments for program evaluation purposes: Many current tools in the inventory have been used for formative assessment—to inform youth so they can improve; and for program implementation purposes, but few were found that have been appropriately validated for program evaluation purposes. Specifically, the field needs tools that are sensitive to program interventions and will detect change over time for individuals or groups on each skill in a program, and which link the individual's or group's performance on each skill to outcomes, depending on the evaluation design.

Implementer inclusion in tool development: Because soft skills measures have historically been developed from the bottom-up through researcher interests, there is often a gap between measures' intentions and everyday realities. Ultimately, if a measure is to serve a practical purpose, practitioners should be included from the outset of measure development. Developers might incorporate practitioner considerations around logistics and practicality of tool administration into tool design as well as leverage partnerships to facilitate tool testing and validation.

Thus, we recommend the following investments to improve the state of youth soft skill measurement.

- A soft skill assessment be developed that draws from the universe of existing tools, is designed specifically for program use, and is appropriate for the age groups of interest. Adaptation might focus first on the high-scoring tools, supplementing as necessary with other relevant items or scales to adequately measure each skill independently, and include age and culturally appropriate language that can be ascertained through cognitive interviews.
- The tool should measure at least the three key cross-cutting skills (positive self-concept, self-control, and higher order thinking skills), using common terminology and definitions developed for this project, that enjoy the strongest evidence across the fields of workforce development, violence prevention, and sexual and reproductive health. Preferably the tool should also include additional skills that enjoy strong support for one or multiple outcome areas: communication, social skills, empathy, goal orientation, positive attitude, and responsibility (see "Key Soft Skills for Cross-Sectoral Youth Outcomes").

- The instrument should be short and easy to administer, translated into languages needed for programs in Latin America, Africa, Middle East, and Asia, and the data resulting from assessments should be easy to analyze and report out.
- The measure might incorporate multiple methods to mitigate the shortcomings of self-report, by comparing with data from other sources. This includes accompanying self-report scales with an observer report method such as observational checklists and/or performance tasks, or a direct assessment, or perhaps a rating from another person, preferably a program staff member. The items should measure frequencies of behaviors that can be reported on by the youth as well as others, which is more objective than endorsing statements. This will involve developing and testing new response scales that accurately report upon and discriminate frequencies of behaviors, particularly at the upper end of the scale.
- Given the need for international adaptation, the measure should be developed and pilot tested in multiple international program contexts. To be most broadly useful for the purpose of evaluating programs' contributions to improving these skills, it should be validated for measuring change over time in those contexts.
- To make tools as relevant as possible to programs, implementers should be included from the outset of tool development.

These steps would make a substantive contribution to the burgeoning field of youth soft skill measurement, enabling greater consistency and efficacy in approaches and, ultimately, much needed comparability across programs. They would also provide an immediate benefit to programs by helping them target assessment and measurement efforts on the most important skills in a cost-effective manner. In addition, they would help youth, program staff, funders, and other key program stakeholders to more concretely capture important aspects of key soft skills while developing programs. Refinement of measures in program implementation and evaluation will help them to understand how and under what conditions they can best foster positive outcomes for youth across sectors. Youth striving to succeed across sectors will benefit from shared understandings of each skill, as well as from improved instruments that are capable of showing them that they have, indeed, improved their skills, which will foster a greater sense of self-efficacy and confidence in their interactions with others and in the workplace.

8. References

Almlund, M., Duckworth, A., Heckman, J.J., and Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, and L. Woßmann (Eds*.), Handbook of the Economics of Education*, Volume 4, pp. 1–181. Amsterdam: Elsevier.

Bartram, D. (2013). Scalar equivalence of OPQ32: Big five profiles of 31 countries. *Journal of Cross-Cultural Psychology*, 44(1), 61–83.

Blades, R., Fauth, B. and Gibb, J. (2012). Measuring Employability Skills: A rapid review to inform development of tools for project evaluation. London: National Children's Bureau.

Card, Noel A. (2016, in press). Methodological Issues in Measuring the Development of Character. University of Connecticut.

Carneiro, P. M., and Heckman, J. J. (2003). *Human capital policy.*

Center for the Economics of Human Development. (2015). Conference on Measuring and Assessing Skills Report. Chicago: University of Chicago.

Connelly, B. S., and Ones D. S. (2010). Another perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, 136(6), 1092–1122. doi:10.1037/a0021212

Deming, D. J. (2015). *The growing importance of social skills in the labor market* (No. w21473). National Bureau of Economic Research.

Duckworth, A. L., and Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237-251.

Educational Testing Service (ETS). (2012). Assessment Methods.

Gates, S., Lippman, L., Shadowen, N., Burke, H., Diener, O., and Malkin, M. (2016). Key Soft Skills for Cross-Sectoral Youth Outcomes. Washington, DC: USAID's YouthPower: Implementation, YouthPower Action.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* 24 (3), 411–482.

Herman, Maureen. "Catholic Relief Services' Youth Build (Jovenes Constructores): Co-Assessment for Soft Skills." 2016. Presentation.

John, O. P., Donahue, E. M., and Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54.* University of California, Berkeley, Institute of Personality and Social Research.
Kane, M.T. (2006). Validation. In *Educational Measurement*, edited by R. Brennan, 4th ed., 17–64. Westport, CT: American Council on Education/Praeger.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., and Borghans, L. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success* (No. w20749). National Bureau of Economic Research.

Kyllonen, P. C., and Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier and D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). Boca Raton, FL: CRC Press.

Kyllonen, P. C. (2015). Designing Tests to Measure Personal Attributes and Noncognitive Skills. In Suzanne Lane, Mark R. Raymond, Thomas M. Haladyna (Eds.,) *Handbook of Test Development*. Abingdon: Routledge.

Lippman, L., Moore, K.A., Guzman, L., Ryberg, R., McIntosh, H., Ramos, M., Caal, S., Carle, A., and Kuhfeld, M. (2014). *Flourishing Children: Defining and Testing Indicators of Positive Development.* Springer Science and Business Media.

Lippman, L.H., Ryberg, R., Carney, R. and Moore, K.A. (2015). Key "Soft Skills" that Foster Youth Workforce Success: Toward a Consensus Across Fields. Washington, DC: USAID, FHI 360, Child Trends. Published through the Workforce Connections project managed by FHI 360 and funded by USAID.

Mischel, W. (2014). *The Marshmallow Test: Mastering self-control*. New York, NY: Little, Brown.

Oh, I. S., Wang, G., and Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, 96(4), 762–773.

Pagel, R.P., Weiss, S., Olaru, G., and Roberts, R. D., (2016). Measuring Youth's Soft Skills Across Cultures: Evidence from the Philippines and Rwanda. Washington, DC: Education Development Center, Professional Examination Services (ProExam), and the Akilah Institute for Women. Published through the Workforce Connections project managed by FHI 360 and funded by USAID.

Roberts, B.W., Walton, K.E., and Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin,* 132(1), 1. DOI:10.1037/0033-2909.132.1.1

Salgado, J. F., and Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies.

Soland, J., Hamilton, L. S., and Stecher, B. M. (2013). Measuring 21st Century Competencies. Global Cities Education Network.

Stecher, B. M., & Hamilton, L. S. (2014). Measuring Hard-to-Measure Student Competencies: A Research and Development Plan. RAND Corporation. Santa Monica, CA.

USAID. Youth in Development: Realizing the Demographic Opportunity. (2012). USAID Policy paper. Washington, D.C.

Wilson-Ahlstrom, A., Yohalem, N., DuBois, D., Ji, P., Hillaker, B., and Weikart, D. P. (2014, January). From soft skills to hard data: Measuring youth program outcomes. Forum for Youth Investment. The Cady-Lee House, 7064 Eastern Avenue NW, Washington, DC 20012-2031.

9. Appendices

Appendix A. Terms Used in Search of Academic Databases and Google Scholar

1) **Skills Search Terms**: "social skills" OR "social intelligence" OR "relationship skills" OR "social competence" OR "conflict management" OR "conflict resolution" OR “social astuteness” OR “social awareness” OR “antisocial behavior” OR “interpersonal skills” OR “social cognitive skills” OR “prosocial norms” OR “communication” OR “active listening” OR “effective communication” OR “effective listening” OR “verbal communication” OR “written communication” OR "non-verbal communication" OR ”higher order thinking skills” OR “decision making” OR “problem solving” OR “critical thinking” OR “deductive reasoning” OR “situational judgment” OR “openness to problem-solving” OR “analytical thinking” OR “reasoning” OR “self-control” OR “self-management” OR “self-discipline” OR “externalizing behaviors” OR “self-regulation” OR “emotional self-regulation” OR “emotional stability” OR "impulsivity” OR “temperance” OR “effortful control” OR “stress management skills” OR “positive self-concept” OR “self-efficacy” OR “self-confidence” OR “self-esteem” OR “positive identity” OR “positive sense of self” OR “positive self-image” OR “self-worth”

2) **Youth Search Terms** – (youth OR adolesc* OR "young adult" OR "young people" OR teen*)

3) **Measurement Search Terms** – “tool” OR “survey” OR “inventory” OR “index” OR “scale” OR “instrument” OR “questionnaire”

Appendix B. Key Informants Interviewed

Maggie Appleton, Educate!
Esther Care, Brookings Institution
Luis Crouch, RTI
Nancy Guerra, University of California-Irvine
Shubha Jayaram, R4D
Tim Kautz, Mathematica
Patrick Kyllonen, ETS
Koji Miyamoto, OECD
Lee Nordstrum, RTI
Ana Maria Munoz-Boudet, World Bank
Rich Roberts, ProExam
Daniel Santos, Ayrton Senna Institute

Appendix C. List of All Tools Reviewed for Initial Screen

1. 12-item Grit Scale
2. 2015 CPS My Voice, My School Student Survey: 6th-12th grade version
3. 21st Century Skills Assessment: Collaboration
4. Academic Skill-Building Program Quality Assessment
5. Achenbach System of Empirically Based Assessments, child behavior checklist
6. Achenbach System of Empirically Based Assessments, clinical interview for children and adolescents
7. Achenbach System of Empirically Based Assessments, youth self-report
8. ACT WorkKeys Listening for Understanding, online
9. ACT WorkKeys Reading for Information, extended
10. ACT WorkKeys Teamwork Assessment, video-based
11. Adolescent Coping Orientation for Problem Experiences Inventory (A-COPE)
12. Adolescent Self-Regulatory Inventory (ASRI)
13. Adolescent Social Self-Efficacy Scale
14. Adult Literacy and Life Skills
15. Adversity Quotient (AQ)
16. Afterschool Program Assessment System
17. Alelo Language and Culture Simulations Language Learning
18. Anchored BFI Tool
19. Arizona Self Sufficiency Matrix
20. Attitudes and Behaviors Survey, Search Institute
21. Balanced Emotional Empathy Scale (BEES)
22. Bar-On EQ-i
23. Barratt Impulsiveness Scale
24. Beck Youth Inventory (BYI)
25. Behavior Assessment System for Children (BASC 2)
26. Behavior Problems Index
27. Berkeley Expressivity Questionnaire
28. Big Five Inventory
29. Bristol Social Adjustment Guide (BSAG)
30. Britain's Key Stage 3 Onscreen ICT Test
31. Buck Institute for Education Presentation Rubric (grades 6 - 8)
32. Buck Institute for Education Presentation Rubric (grades 9 - 12)
33. California Health Kids Survey (CHKS): Supplementary Module
34. California Healthy Kids Survey (CHKS): Core Module, high school
35. California Healthy Kids Survey (CHKS): Core Module, middle school
36. California Healthy Kids Survey (CHKS): GRAM safety module (gang risk)
37. California Healthy Kids Survey (CHKS): Resilience and Youth Development Module
38. California Healthy Kids Survey (CHKS): Social Emotional Health Model
39. Camper Growth Index-Camper (CGI-C)
40. CARALOC Pupil Sample Questionnaire
41. Casey Life Skills Assessment
42. Child and Adolescent Wellness Scale (CAWS)
43. Child Behavior Checklist (CBCL)
44. Child Perceived Self-Efficacy Scale

45. Child Trauma Screening Questionnaire
46. Children and Adolescent Needs and Strengths
47. Children's Hope Scale
48. Chinese Positive Youth Development Scale (CPYDS)
49. Civic Engagement VALUE Rubric (Association of American Colleges and Universities)
50. Collaboration Rubric, Catalina Foothills Schools District
51. Collegiate Learning Assessment
52. Communication Rubric, Catalina Foothills Schools District
53. Communication Scale, Youth Life Skills Evaluation Project at Penn State
54. Communities that Care Survey (CTCYS)
55. Competency-Based Education Assessment, P-21 Collegiate Learning Assessment
56. Competent Speaker Speech Evaluation Form (CSSEF)
57. Comprehensive Assessment of Team Member Effectiveness (CATME)
58. Condom Self-Efficacy Scale
59. Conflict in Adolescent Dating Relationships Inventory
60. Conflict Resolution—Individual Protective Factors
61. Conscientiousness Facets Tools
62. Conversational Skills Rating Scale (CSRS)
63. Coping Scale for Children and Youth
64. Core Competencies (Student Questionnaire Based on the 5 C's)
65. Cornell Critical Thinking Tests
66. Creative Thinking VALUE Rubric (Association of American Colleges and Universities)
67. Creativity and Innovation Rubric, Catalina Foothills Schools District
68. Critical Thinking, adapted from Youth Engagement, Attitudes, and Knowledge (YEAK) Survey
69. Critical Thinking and Problem Solving Rubric, (CSFD) Catalina Foothills Schools District
70. Critical Thinking Assessments, P-21 Collegiate Learning Assessments
71. Critical Thinking in Everyday Life, Youth Life Skills Evaluation project at Penn State
72. Critical Thinking Performance Assessment (CWRA+), Council for Aid to Education
73. Critical Thinking Scale, Youth Life Skills Evaluation Project at Penn State
74. Critical Thinking Value Rubric (Association of American Colleges and Universities)
75. DAP+ (Developmental Assets Profile plus Workforce Readiness Skills)
76. Developmental Indicators for the Assessment of Learning (DIAL) Assessment
77. Devereux Student Strengths Assessment, DESSA scales and associated items
78. Differentiation of Self Inventory (DSI)
79. Diligence and Reliability Scale, Child Trends
80. Early Childhood Hope Inventory
81. East African Youth Creativity Scale
82. East African Youth GRIT Scale
83. East African Youth Public Speaking Scale
84. East African Youth Risk-taking Behavior Scale
85. East African Youth Self-Confidence Scale
86. East African Youth Self-Efficacy Scale
87. East African Youth Self-Esteem Scale
88. Ecological Multi-User Virtual Environment (EcoMUVE)
89. EdLeader21 Communication Rubrics
90. Educational Longitudinal Study Student Questionnaire, base year and follow-up

91. Emotion Regulation Questionnaire
92. Employability Competency System—Pre-employment Work Maturity Checklist
93. Employee Attitude Inventory (EAI)
94. Employer Survey Research Report, Learning and Skills Network
95. Equipped for the Future (EFF) Assessment Prototype
96. Equipped for the Future Performance Standards
97. Ethical Reasoning VALUE Rubric (Association of American Colleges and Universities)
98. ETS WorkFORCE® Assessment for Job Fit
99. Eyberg Child Behavior Inventory
100. Fast Track, Conduct Problems Prevention Research Group (CPPRG)
101. Feelings and Emotions Scale (PANAS-C)
102. Flourishing Children Study, Adolescent Scale
103. Flourishing Children Study, Parent Scale
104. Foundations and Skills for Lifelong Learning VALUE Rubric (Association of American Colleges and Universities)
105. General Self-Efficacy Scale
106. Global Competitiveness Assessment Tool
107. Global Learning VALUE Rubric, Association of American Colleges and Universities
108. Global Youth Well-Being Index, CSIS
109. Hall of Heroes
110. Health & Daily Living Form (HDLF)
111. Hogan Personality Inventory Reliability Scale
112. Holistic Student Assessment
113. Index of Self-Esteem, Walter Hudson
114. Individual Protective Factors Index
115. Information Literacy VALUE Rubric, Association of American Colleges and Universities
116. Initiative Taking Scale, Child Trends
117. Inquiry and Analysis VALUE Rubric, Association of American Colleges and Universities
118. Integrative Learning VALUE Rubric, Association of American Colleges and Universities
119. Intercultural Competence and Knowledge VALUE Rubric, Association of American Colleges and Universities
120. International Men and Gender Equality Survey
121. International Youth Development Survey (IYDS)
122. Inventory of Parent and Peer Attachment
123. Jackson Personality Inventory
124. Jamaica Youth Survey
125. Jesness Inventory (revised)
126. Job Performance Personality Inventory
127. Jovenes Constructores Competencies Self-Evaluation
128. Knack
129. LAWSEQ Pupil Sample Questionnaire
130. Learn, Earn, and Save
131. Locus of Control Survey
132. Massachusetts Work-based Learning Plan
133. Matson Evaluation of Social Skills with Youngsters (MESSY)
134. Mayer-Salovey-Caruso Emotional Intelligence Test

135. Measuring Elementary School Students' Social and Emotional Skills: Teacher's Survey, Child Trends
136. Michigan Adolescent and Adult Life Transitions Survey
137. Middle Years Development Instrument, grade 4
138. Middle Years Development Instrument, grade 7
139. Mission Skills Assessment (MSA), Collaboration
140. Motivated Strategies for Learning Questionnaire (MSLQ)
141. Motivated Strategies for Learning Questionnaire (MSLQ): Subscale Self-Regulated Learning Strategies
142. Motivational and Self-Regulated Learning Components of Classroom Academic Performance
143. Multidimensional Scale of Perceived Social Support
144. Multidimensional Self-Concept Scale
145. National Assessment of Educational Progress (NAEP) Reading & Writing Tests
146. National Career Readiness Certificate
147. National Longitudinal Survey of Adolescent Health
148. NEO Personality Inventory
149. New Self-Efficacy Scale (new GSE scale (NGSE))
150. Oral Communication VALUE Rubric (Association of American Colleges and Universities)
151. Orphans and Vulnerable Children CWB tool
152. Participant Work Readiness Evaluation
153. Partner Communication Scale
154. Partnership for Assessment of Readiness for College and Careers (PARCC)
155. Passport to Success: Retrospective Completion Survey
156. Passport to Success: Trainer Observation Tool
157. Personal Leadership Inventory, all versions
158. Personal Potential Index (PPI)
159. Personal Resilience Inventory
160. Personnel Selection Inventory
161. Piers-Harris Children's Self-Concept Scale
162. PISA Problem Solving Computer Test
163. PISA Problem Solving Experiences, section F of PISA student questionnaire
164. Political Skills Inventory
165. Positive Youth Development Student Questionnaire
166. Practical & Soft Skills, Educate!
167. Pre-Employment Work Maturity Checklists
168. Problem Solving, (adapts the Problem Solving Scale from the Work Keys)
169. Problem Solving VALUE Rubric (Association of American Colleges and Universities)
170. Profile of Student Life—Attitudes and Behaviors
171. Profile of the Young Entrepreneur in the West Bank
172. Program for the International Assessment of Adult Competencies (PIAAC) Background Questionnaire, Module F Skills Used at Work
173. Program for the International Assessment of Adult Competencies (PIAAC) Problem Solving in Technology Rich Environments
174. Prosocial Personality Battery
175. Psychological Maturity Inventory
176. Quality of Hire Talent Scorecard

177. Quality Rubric, Citizen Schools
178. Quantitative Literacy VALUE Rubric (Association of American Colleges and Universities)
179. Questionnaire on Self-Regulation
180. REACH Survey
181. Reading VALUE Rubric (Association of American Colleges and Universities)
182. Resilience & Youth Development Module – full version (protective factors)
183. Resilience Scale Survey
184. Responses to Stress Questionnaire (RSQ)
185. Rosenberg Self-Esteem Scale
186. Rotter's Locus of Control Scale
187. Safety & Violence Module - High School Survey
188. Safety & Violence Module - Middle School Survey
189. Satisfaction with Life Scale, adapted for children
190. School Engagement Scale—Behavioral, Emotional and Cognitive Engagement
191. Secondary Skills Assessment Tool (SSAT)
192. Self-Esteem Questionnaire, DuBois, D. L., Felner, R. D., Brand, S., & Phillips, R. C. (1996)
193. Self-Perception Profile for Children (SPPC)
194. Self-Regulation Questionnaire, Brown, Miller, & Lawendowski (1999)
195. Senna 1.0
196. Senna 2.0
197. Sexual Behavior Module - High School survey
198. Sexual Behavior Module - Middle School survey
199. SimScientists
200. Singapore Group Project Portfolio
201. Smarter Balanced Assessment
202. Social and Person Competence Scales
203. Social Competence Scale for Teenagers
204. Social Competence Tool, Child Trends
205. Social Emotional Health Survey System
206. Social Emotional Skills – School Quality Improvement Index (SQII)
207. Social Skills Improvement System (SSIS)
208. Social Skills Inventory, Ronald E. Riggio (1989, 2002)
209. Speaking and Listening, P21-Collegiate Learning Assessment (CLA)
210. Step it Up 2 Thrive – Life Skills, 12th
211. Step It Up 2 Thrive – Life Skills, 9th – 11th
212. Step it Up 2 Thrive – Social Skills
213. Step Skills Measurement Surveys
214. Strengths and Difficulties Questionnaire (SDQ), 11-17 (parent or teacher)
215. Strengths and Difficulties Questionnaire (SDQ), 18+ (self-report)
216. Student Engagement Instrument
217. Student Leadership Practices Inventory
218. Survey of Academic and Youth Outcomes Youth Survey (SAYO)
219. Survey of Elementary School Students Social and Emotional Skills, student survey, Child Trends
220. Survey of Elementary School Students Social and Emotional Skills, teacher survey, Child Trends

221. Teamwork VALUE Rubric (Association of American Colleges and Universities)
222. Ten Item Personality Index (TIPI)
223. Tennessee Self-Concept Scale
224. Torrance Test of Creative Thinking
225. Trait Emotional Intelligence Questionnaire
226. Trustworthiness and Integrity Scale, Child Trends
227. UKCES Employer Skills Survey
228. Wakefield Resilience Framework
229. Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS), 14-item scale
230. Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS), 7-item scale
231. Washington Healthy Youth Survey
232. WIA Work Readiness tool for Youth, U.S. Department of Labor
233. Work Star, Outcomes Star
234. Workforce Skills Certification System
235. Working (Assessing Skills, Habits, and Style) Scales
236. World Bank Skills towards Employment and Productivity (STEP) Skills Measurement
237. Written Communication VALUE Rubric (Association of American Colleges and Universities)
238. Young Risk Behavior Study
239. Youth Connections Scale, A. Semanchin-Jones & T. LaLiberte (2012)
240. Youth Empowerment Scale—Mental Health
241. Youth Program Quality Assessment, 5-12 year olds
242. Youth Program Quality Assessment, 9-18 year olds
243. Youth Programs and Strengths Survey
244. Youth Psychopathic Traits Inventory

Appendix D. Key Sources of Measurement Tools

Afterschool Outcome Measures Online Toolbox. (n.d.). Retrieved from http://afterschooloutcomes.org/.

Association of American Colleges & Universities. (2015). Retrieved from https://www.aacu.org/value-rubrics.

Balcar, J. (2011). *Transferability of Skills across Economic Sectors: Role and Importance for Employment at European Level*. Publications Office of the European Union.

Banyard, V. L. "Interpersonal Violence in Adolescence: Ecological Correlates of Self-Reported Perpetration." *Journal of Interpersonal Violence* 21.10 (2006): pp. 1314-332.

Blades, R., Fauth, B., & Gibb, J. (2012). Measuring employability skills: A rapid review to inform development of tools for project evaluation. London: National Children's Bureau.

CASEL. Core SEL Competencies. (n.d.). Retrieved from https://www.casel.org/core-competencies/.

Compendium for School Climate Surveys. (n.d.). Retrieved from http://safesupportiveschools.gov.

Cauffman, E., Kimonis, E. R., Dmitrieva, J., & Monahan, K. C. (2009). A multimethod assessment of juvenile psychopathy: Comparing the predictive utility of the PCL:YV, YPI, and NEO PRI. *Psychological Assessment*, 21(4). doi:10.1037/a0017367.

Catalina Foothils School District. Resources for Deep Learning. (n.d.). Retrieved from http://www.cfsd16.org/index.php/academics/resources-for-deep-learning.

Child Trends. (2014). Measuring Elementary School Students' Social and Emotional Skills: Providing Educators with Tools to Measure and Monitor Social and Emotional Skills that Lead to Academic Success. Bethesda: MD.

Child Trends. Positive Indicators Project. Retrieved from http://www.childtrends.org/positive-indicators-project/. Bethesda: MD.

CLA Supporting Materials. (n.d.). Retrieved from http://cae.org/participating-institutions/cla-references/cla-supporting-materials/.

Cooper, L.M.,Wood, P.K., Orcutt, H.K., Albino, A. (2003). Personality and the predisposition to engage in risky or problem behaviors during adolescence. *Journal of Personality and Social Psychology*, 84(2), 390-410.

Core Districts. (2016). "Social-Emotional Learning." Retrieved from http://coredistricts.org/our-work/social-emotional-learning/.

Department of Education. Employability Skills Framework. Retrieved from http://cte.ed.gov/employabilityskills/index.php/assessment/custom_worksheet_step1.

Duckworth, A.L, & Quinn, P.D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment, 91,* 166-174. Retrieved from http://www.sas.upenn.edu/~duckwort/images/Duckworth%20and%20Quinn.pdf

Ennis, R. H. (n.d.). Critical Thinking. In *The Palgrave Handbook of Critical Thinking in Higher Education*. doi:10.1057/9781137378057.0005.

Gross, J.J., & John, O.P. (1997). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, 72, 435-448.

Gross, J.J., & John, O.P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85, 348-362.

Gutman, L. M., & Schoon, I. (2003). The impact of non-cognitive skills on outcomes for young people: Literature review. London: Institute of Education, University of London.

Haggerty, K., Elgin, J., & Woolley, A. (2011). Social-emotional learning assessment measures for middle school youth. Social Development Research Group. University of Washington: Raikes Foundation.

Hahn, Andy, Susan Lanspery and Tom Leavitt. (2006). Measuring Outcomes in Programs Designed to Help Young People Acquire Life Skills. The Heller School for Social Policy and Management Center for Youth and Communities, Brandeis University.

International Personality Item Pool surveys. (n.d.). Retrieved from http://ipip.ori.org/.

Kyllonen, P. C. (2013). Soft skills for the workplace. *Change: The Magazine of Higher Learning*, *45*(6), 16-23.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2009). Personality, motivation, and college readiness: A prospectus for assessment and development. Princeton, NJ: Educational Testing Service.

Lerner, R. M., & Lerner, J. V. (2013). The positive development of youth: Comprehensive findings from the 4-H study of positive youth development. Chevy Chase, MD: National 4-H Council.

Marsh, H. W., Richards, G. E., & Barnes, J. (1986). Multidimensional Self-Concepts: A Long-Term Follow-Up of the Effect of Participation in an Outward Bound Program. *Personality and Social Psychology Bulletin*, 12(4). doi:10.1177/0146167286124011.

McNeil, B., Reeder, N., & Rich, J. (2012). A framework of outcomes for young people. *Young Foundation*.

National Institute on Out-of-School Time. (n.d.). Retrieved from http://niost.org/.

Olenik, Christina, Zdrojewski, N., and Bhattacharya, S. “Scan and Review of Youth Development Tools.” Washington, DC: USAID.

Pagel, R.P., Weiss, S., Olaru, G., and Roberts, R. D., (2016). Measuring Youth’s Soft Skills Across Cultures: Evidence from the Philippines and Rwanda. Washington, DC: Education Development Center, Professional Examination Services (ProExam), and the Akilah Institute for

Women. Published through the Workforce Connections project managed by FHI 360 and funded by USAID.

Partnership for 21st Century Skills. (2007). "21st Century Skills Assessment: A Partnership for 21st Century Skills e-paper."

Platt, Gary (2008). The Hard Facts About soft skills Measurement. *Training Journal*, August, pp. 53-56.

Simmons, C. (2012). *Tools for strengths-based assessment and evaluation*. Springer Publishing Company.

Smith, C., McGovern, G., Larson, R., Hillaker, B., & Peck, S. C. (2016). Preparing Youth to Thrive: Promising Practices for Social and Emotional Learning. Ypsilanti, MI: David P. Weikart Center for Youth Program Quality at the Forum for Youth Investment.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). Measuring 21st Century Competencies: Guidance for Eduators.

Stecher, B. M., & Hamilton, L. S. (2014). Measuring Hard-to-Measure Student Competencies: A Research and Development Plan. Research Report. RAND Corporation. Santa Monica, CA.

STEM Learning and Research Center. (n.d.). Retrieved from http://stelar.edc.org/instruments/motivated-strategies-learning-questionnaire-mslq.

Surveys of CPS Schools. (n.d.). Retrieved from https://consortium.uchicago.edu/surveys.

The Skills Library. (n.d.). Retrieved from http://www.skillslibrary.com/wbl.htm

University of Minnesota CYFAR Tool Bank. (n.d.). Retrieved from https://www.cyfar.org.

Wagner-Rundell, Nicole. (2016). Social and Emotional Outcome Measurement: A Toolbox. Child Trends.

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

Wadsworth, M. E., & Compas, B. E. (2002). Coping with Family Conflict and Economic Strain: The Adolescent Perspective. *Journal of Research on Adolescence*, 12(2), 243-274. doi:10.1111/1532-7795.00033.

Windle, G., Bennett, K. M., & Noyes, J. (2011). A methodological review of resilience measurement scales. *Health and quality of life outcomes*, *9*(1), 8.

U.S. Agency for International Development
1300 Pennsylvania Avenue, NW Washington, DC 20523
Tel: (202) 712-0000
Fax: (202) 216-3524

www.usaid.gov