BEST PRACTICES FOR MEASURING AND REPORTING USAID'S YOUTH SKILLS INDICATORS

USAID's five youth skills indicators (see Figure 1), as discussed in USAID's Office of Education's How-To Note, Measuring Skills for Youth Workforce Development, call for appropriate assessment planning and implementation, sampling, and analysis and reporting. This document presents best practices for planning and implementing assessments, selecting participants and sampling, and analyzing and reporting on the five indicators. While many of these best practices are universal to rigorous measurement and analysis, they are contextualized here for USAID's youth skills indicators.

FIGURE 1.

SKILL	PREVIOUS INDICATOR	NEW INDICATOR
Standard Foreign Assistance Indicators (required as applicable)		
Soft Skills	EG. 6-2 Number of individuals with improved skills following completion of USG-assisted workforce development programs	EG. 6-13 Percent of individuals with improved soft skills following participation in USG-assisted workforce development programs
Reading	N/A	ES.1-54 Percent of individuals with improved reading skills following participation in USG-assisted youth programs
Supplemental Indicators (encouraged as applicable)		
Math	N/A	Percent of individuals with improved math skills following participation in USG-assisted programs
Digital Literacy	N/A	Percent of individuals with improved digital literacy skills following participation in USG-assisted programs
Technical/ Vocational/ Professional Skills	N/A	Percent of individuals who pass a context-relevant assessment in a technical, vocational, or professional skill set following participation in USG-assisted programs

Planning and Implementing Assessments

All of USAID's Office of Education's youth indicators call for appropriate assessment validity, reliability, and context-relevance. Implementers should follow the below best practices for planning and implementing assessments to ensure they meet these standards, as is feasible and reasonable.

Validate Assessments in the Context and for the Target Population. The chosen assessment method(s) should be validated and reliable for the particular population of participants in that particular context or a similar one. When ensuring an assessment is appropriate for the "context," implementers should consider the following exemplary—though not exhaustive—characteristics:

- Language and Culture. Due to language and cultural differences, many skills assessments cannot simply be directly translated and assumed to be relevant for new contexts—certain scenarios or cultural reference points may be familiar in one place but not another, or certain skills or applications of those skills may be desirable in one labor market context but not in another.
- Age. The notion of context-relevance is also related to age-appropriateness: assessment items should not reference vocabulary or experiences that are unfamiliar for the target population due to their age or any other factor.
- Market Context. With reference to the indicator for technical, vocational, and professional skill sets, "context-relevance" also takes on additional, market-related meanings.
- Reading Level. Written assessments of soft skills, math skills, digital literacy, and technical, vocational, and professional skills should also be set at the appropriate reading level for beneficiaries in general and should not significantly hinder the performance of certain beneficiaries who may have a lower reading level than that of their peers.
- Cognitive Load coupled with Poverty Levels. Item and response structure should not be so complex that they represent an unreasonable cognitive load that interferes with the performance of beneficiaries who have less formal education, lower concentration skills, or whose poverty may affect cognitive functioning.[1]

If implementers must develop their own assessments for any youth skills indicators, the possibility of gender or cultural bias of certain items or the assessment as a whole must be analyzed after piloting by examining the scores of these groups in comparison to one another to check for systematic gender- or culture-related patterns of performance on specific items or scales.

Use criterion-referenced assessments—where scores are either automatically calculated or pegged to known standards expressed in a rubric form—to avoid score inflation, a circumstance in which an increase in an individual's test score does not reflect an increase in that individuals' understanding of the content. Norm-referenced assessments, in which scores are calculated in relation to the performance of other participants, are not permitted for the purposes of these indicators; the only exception may be for technical, vocational, or professional assessments that are adopted from an existing assessment institution and which may use norm-referencing to assign pass rates.

Limit access to the instrument before it is administered to reduce opportunities for corruption—when the test creator or administrator provides answers to the questions on an assessment— and cheating. When the same assessment will be used at both baseline and endline, implementers need to ensure instructors know that improved assessment results will carry no rewards for either participants or instructors. Whenever activities tie reward or punishment to skills assessments, even if it is simply in the form of recognition or praise, instructors may have the incentive to cheat or otherwise assist participants to attain higher scores.

[1] Mani, A., Mullainathan, S., Shafir, E., Zhao, J. (30 August 2013). "Poverty Impedes Cognitive Function." *Science*. Vol 341. https://scholar.harvard.edu/files/sendhil/files/976.full_.pdf

Consider the optimal timing of the assessments. The intent of a pre-test or pre-assessment is to capture the state of participants' skills prior to any organized efforts to raise their skill level. Capturing pre-test data as early as possible typically results in the ability to demonstrate greater program effects over time. However, there may be cases when it is advisable to administer the pre-test slightly after the beginning of the program in order to introduce participants to the programming before they can accurately assess their own skill level. Implementers will therefore need to use judgment and experience to determine the best moment within their program to administer the pre-test.

Post-tests should, for the purposes of these indicators, be administered immediately or close to the completion of organized learning activities that specifically concern the skill area to be assessed. Even if a tracer study is conducted, implementers should report on these indicators using the assessment data collected just after the end of the training or direct intervention, rather than assessment data collected later on.

Reduce bias by using different forms at baseline and endline. Where possible, it is good practice to have two different sets of items that have been fully equated so that participants do not respond to the same questions at the beginning and end of the program. However, USAID acknowledges that it is often difficult, time-consuming, and expensive to ensure such equated test forms, particularly in newer assessment areas such as soft skills. Where it is necessary to use the same pre-test and post-test forms, implementers can take measures to reduce biased effects such as randomizing the order of items on the post-test.

Ensure assessments measure individual-level improvement in skills. Four out of the five youth skills indicators require measuring improvements in the skills of individuals from one time to another. Yet, many existing assessments included in the Measuring Skills for Youth Workforce Development How-To Note's tables of example tools have only been explicitly validated for comparing group-level results or trends, meaning that the available published documentation discusses changes in the average scores of a group at two points in time, or comparisons of the average scores of different groups at the same time, rather than individual-level pathways of score improvements. Thus, many of the referenced assessments in the How-To Note do still require additional work to show that they are valid for measuring individual-level change.

Selecting Participants and Sample Design

While there are many different and accurate ways to determine which participants will participate in a study or to design a sample, characteristics of USAID's youth skills indicators limit the options available for sample design. In this section, USAID presents key expectations for selecting participants and sampling for youth skills indicators.

Examine all participants, not just completers. Most activities experience attrition or drop-out. Even individuals who drop out, however, may have received some benefit from the program during the time that they were able to be engaged. In addition, USAID is interested in gathering data on the population an activity intended to serve or deliver services to (known as an "intent to treat" focus), since this approach better reflects the potential real effects and costs of large-scale adoption of a program, such as by public government institutions. For these reasons, the indicators formulated here deliberately refer to those who have "participated" in programming, rather than those who have "completed" a given course of instruction or set of interventions.

Define "Participants." The sampling frame for all of these skills indicators is the population of all participants. Participants are those who are still in the program at the end of the program's enrollment phase for that class or session. Because USAID-funded programs are so diverse and different, each program must determine for itself what constitutes the enrollment phase. For example, an enrollment phase might be described as the first few days of programming, after which the activity is accepting no new participants. The definition of enrollment phase that the program decides upon must be included in the activity's monitoring, evaluation and learning (MEL) plan.

Sample to ensure representativeness. Indicator reports that rely on a sample of participants rather than a census to report results should construct their pre-test sample to ensure representation of characteristics that appear as indicator disaggregates (sex is a required disaggregate for USAID, and in some cases, disability and conflict or crisis-status are as well) or are important for understanding differences in outcomes (e.g., geography, language, etc.). This implies use of a stratified sampling method, which begins with grouping participants by the characteristics required for that indicator's disaggregation categories. Sampling should still be done randomly within these characteristics (producing a set of "stratified random samples") to ensure that sampled participant characteristics provide an unbiased of the estimate the population characteristics.

Trace all (sampled) participants for post-test, as feasible. In order to determine the skill gains of all participants, rather than just those who completed the program, it is necessary to make the attempt to trace and assess even those who dropped out of the program. This means that implementers must collect adequate contact information for all participants included in their pre-test sample (including, for example, telephone numbers of family members or neighbors as backup) in order to be able to seek out participants who are not present on the day of program completion and post-test assessment. This is not a requirement for implementers to engage in extensive new efforts to re-engage participants who have dropped out, but rather simply to make a reasonable attempt to follow up with the entire original sample regardless of their current status, much as they would for a youth who completed the entire program but just missed the final assessment.

Understandably, programs may not be able to locate and assess all of those who are absent at the final assessment time. In this case, the indicator will be based on a comparison of the individual pre-test and post-test scores of each person that was assessed at both times.

Weight, as appropriate, and extrapolate data onto the activity population. In order to extrapolate findings from the sample to the activity's population of participants, it is important to know whether the final sample still reflects the whole population of participants on all major characteristics. If it does not, adjustments need to be made. At a minimum, implementers should determine whether the final sample reflects the whole population of participants with respect to the major disaggregation criteria or stratifications, and the percent of participants who completed the whole program. Implementers may determine other important characteristics to compare as well. If the final sample does not reflect the whole population of participants, then the sample should be appropriately weighted to reflect the characteristics of the whole population. It is possible to weight the data on multiple factors at once to obtain more accurate results.[2]

[2] For more more detailed guidance see, for example, http://www.applied-survey-methods.com/weight.html

Analyzing Data and Reporting Indicators

The indicators for soft skills, reading skills, math skills, and digital literacy skills have similar analysis and reporting requirements because they all call for the use of a longitudinal pre/post assessment design,[3] and they allow implementers to choose whether to report on the indicators based on a census of participants or a representative sample extrapolated to the activity population.[4] Best practices for analyzing data and reporting on these indicators is addressed below.

Define improvement according to context, the activity, and the tool. Wherever possible, define "improvement" according to the movement from one to another benchmark or level already established in validated assessments. In the case of reading and math skills, rubrics describing meaningfully distinct skill levels that are internationally accepted have been provided to assist with benchmarking assessments that do not have their own internally defined levels. In the case of soft skills and digital literacy, USAID suggests that meaningful improvement in an individual's score be determined and defined by implementers themselves. For all skills, the targeted level or degree of improvement will differ depending on context, characteristics of the activity, and the assessment tool.

Disaggregate data as required by the PIRS. Each of the indicators should be calculated for the total population of program participants, with both numerators and denominators reported. When a sample is used, the numerator should be extrapolated to the population of program participants. Additionally, data should be disaggregated by sex, age band, disability status, and individuals affected by crisis or conflict. Some of these disaggregations require representative sampling strategies as well; when this is the case, it is noted in the PIRS.

While outcomes will not be disaggregated by the type or intensity/dosage of the intervention, implementing partners are encouraged to provide ample detail on these factors in their quarterly, annual, or evaluation reports in order to enable a more in-depth analysis of the data collected through these indicators.

Avoid double counting within an indicator. In preparing for data analysis, each individual's results should be counted only once per indicator, regardless of the number of training opportunities in which the individual participated. For example, when an individual has participated in multiple digital literacy skills trainings, endline assessment should occur only at the end of the individual's participation in all digital skills programming. The same individual can, however, be counted toward multiple different skills indicators, if they improved on multiple different types of skills (for example, digital literacy, soft skills, and reading skills).

[3] The indicator for technical, vocational, and professional skills focuses only on a post-course assessment measuring the attainment of a certain industry-relevant skill level, since this is the manner in which these skills are most typically assessed by existing systems; see indicator explanation for further details.

[4] Primarily for equity reasons, the indicator for technical, vocational, and professional skills requires direct measurement of all eligible individuals (a "census-based" rather than sample-based approach); see indicator explanation for further details.