USAID
FROM THE AMERICAN PEOPLE

YOUTHPOWER ACTION

YOUTHPOWER ACTION

Measuring Soft Skills Among Youth and Young Adults:
Validation of a New Instrument

TECHNICAL REPORT

February 2020

This publication was made possible by the support of the American People through the U.S. Agency for International Development(USAID), with support from the U.S. President's Emergency Plan for AIDS Relief, under task order contract number AID-OAA-TO-15-00003, YouthPower Action under IDIQ contract number AID-OAA-I-15-00009, YouthPower: Implementation. The contents of this publication 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.

YOUTHPOWER ACTION

Measuring Soft Skills Among Youth and Young Adults:
Validation of a New Instrument

TECHNICAL REPORT

This publication was produced for review by the United States Agency for International Development. It was prepared by Carina Omoeva, Sarah Gates, Nina Cunha, Andres Martinez, and Holly Burke, and FHI 360.

Acknowledgements

This report was prepared by the FHI 360 YouthPower Action | Measurement of Soft Skills Team: Carina Omoeva, Sarah Gates, Nina Cunha, Andres Martinez, and Holly Burke, under the overall guidance of Kristin Brady, the YouthPower Action Project Director. The authors would like to thank USAID E3 for their oversight and support throughout the study, as well as review and feedback on earlier versions of this report: Nancy Taggart, Olga Merchan, Rebecca Pagel, Laurette Cucuzza, and the YouthPower Action COR Elizabeth Berard. The team is indebted to Pat Kyllonen, for his continuous expert input, review and feedback on multiple versions of the measurement instrument, and critical assistance in the analytical stage, particularly on measurement invariance analysis and anchoring vignette integration. The team is also grateful to Laura Lippman for her formative work that informed the instrument development process, including earlier research that led to the prioritization of soft skills and review of the measurement landscape for soft skill, as well as her early input into the tool construction.

Finally, YouthPower Action is grateful to the members of the Soft Skills Measurement Technical Advisory Group who provided early input into on the development and study protocol for the validation of this instrument[1]:

- Richard Lerner, Tufts University
- Andrea Lozano, Save the Children
- Karen Moore, Mastercard Foundation
- Lee Nordstrum, RTI International
- Daniel Santos, Aryton Senna Institute, University of Sao Paolo
- Teresa Wallace, World Vision

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: Omoeva, C., Gates, S., Cunha, N., Martinez, A., and Burke, H. (2020). Measuring Soft Skills Among Youth and Young Adults: Validation of a New Instrument. Washington, DC: USAID's YouthPower: Implementation, YouthPower Action.

[1] Affiliations indicated reflect individuals' institutional affiliations during the period in which the Technical Advisory Group was convened (2017).

Table of Contents

Table of Contents 2
Acknowledgements 2
Acronyms 5
Executive Summary 6

Introduction 11
Background 11
Methodology 12
Instrument Design 13
Field Testing 14
Data Analysis 16
Program Staff Assessment 25
Analysis of Item Order 27
Analysis of Enumerator Characteristics 27
Conclusion 30
Recommendations 31
References 34

Appendix Materials 37
Appendix 1. Five-Factor versus Four-Factor Structure of Soft Skills Tool 37
Appendix 2. Scale Sample Composition and Descriptive Statistics 44
Appendix 3. Additional Data from Exploratory Factor Analysis 59
Appendix 4. Background on Measurement Invariance Analysis Process 61
Appendix 5. Background on DIF Analysis 63
Appendix 6. Additional Data from Anchoring Vignette Analysis 65
Appendix 7. Additional Information on Program Staff Tool Analysis 66

Acronyms

AV	Anchoring Vignettes
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CRS	Center for Social Research
DIF	Differential Item Functioning
EFA	Exploratory Factor Analysis
HOTS	Higher Order Thinking Skills
M&E	Monitoring and Evaluation
MI	Measurement Invariance
MG- CFA	Multiple Group Confirmatory Factor Analysis
OECD	Organization for Economic Cooperation and Development
RH	Reproductive Health
RMSEA	Root Mean Square Error of Approximation
SES	Socio Economic Status
SRMR	Standardized Root Mean Square Residual
TLI	Tucker-Lewis Index
US	United States
USG	United States Government
USAID	United States Agency for International Development

Executive Summary

The YouthPower Action Youth Soft Skills Assessment is a measurement tool for youth soft skills, developed and validated for administration with youth in lower resource environments. The target group for the measurement tool is youth aged 15-19 years, both in and out-of-school, who are typically the beneficiaries of USAID-sponsored youth programs. The design and testing initially included two instruments: a youth self-assessment, as well as a third-party assessment, intended for use with youth program mentors or facilitators. This report describes the process of instrument development, as well as presents the results of the testing and validation at two sites.

The Youth Soft Skills Assessment is grounded in an extensive review of literature and draws from a repository of close to 300 instruments intended to measure cross-sectoral skills. This instrument seeks to measure key skills predictive of positive youth outcomes: workforce success, violence prevention, and reproductive health, as well as be: a) targeted to youth ages 15-19; b) low cost and easy to administer, c) sensitive to change over time, and d) consistent across cultural contexts. The instrument development consisted of the following stages:

1. Initial instrument design, based on a repository of existing tools and adaptation;
2. Field testing in Uganda and Guatemala, consisting of qualitative cognitive interviews and field administration of the tool in both countries at two points in time;
3. Revisions to item content and wording, as well as response options; exclusion of poorly performing items;
4. Validation analysis, including scale structure (exploratory and confirmatory factor analysis, reliability analysis, predictive validity, analysis of change over time, and cross-cultural comparability).

Instrument Design. Drawing from an item bank constructed from the repository of existing assessments, the youth assessment includes items that measure positive self-concept, self-control, higher-order thinking skills, and social and communication skills. In subsequent field testing and analysis, these scales were revised to include: **positive self-concept**, **negative self-concept**, **higher-order thinking skills**, and **social and communication skills**. Items were initially drafted as questions about behavioral frequency (e.g. "How often do you…?"), with a 6-point behavioral frequency scale, from "Almost Never or Never" to "Almost Always or Always". Following the first round of testing, items were revised to statements about self or others, with a 4-point endorsement scale for response options, from "Strongly Disagree" to "Strongly Agree". A series of *anchoring vignettes (AVs),* intended to circumvent cultural response bias, were included in the youth self-report component of the tool. For validation purposes, the instrument also includes a module of socio-economic status proxies, a module on disability status, and a series of items measuring key outcomes of interest, including employment, reproductive health, and violence and conflict behaviors.

Program Staff Tool. The youth self-assessment was initially accompanied by a third-party assessment module, to be administered by youth mentors or facilitators. This was in response to findings from the measurement tool review and feedback from a group of technical experts that many existing soft skills tools suffer from self-report bias. However, in subsequent testing and validation the program staff module was found to have limited value and little reliability.

Field Testing. The youth self-assessment and program staff module were tested in two different sites: in Uganda with the Educate! Program, a non-US Government program for youth in secondary schools, and in Guatemala with the Proyecto Puentes, a USAID funded program implemented by World Vision in the Western Highlands of Guatemala that delivers life skills and technical and vocational training among 15-24-year-olds, both in-school and out-of-school. As a first step, **qualitative cognitive interviews**

served to assess the content validity of the items, response options, and overall structure of the tool; and processes for electronic data capture and analysis. Cognitive testing informed the final set of response options, item wording, content of some items, and exclusion of some items.

Data collection took place in two countries. In Uganda, 1,089 youth participants of *Educate!,* a youth entrepreneurship program in secondary schools across the country, were surveyed at baseline in March 2018. Fifty-nine schools were sampled for the study, with 19 youth per school, including both *Educate!* scholars and age-eligible non-scholars. In Guatemala, 794 youth participants of the USAID program Proyecto Puentes, a USAID program, were tested at baseline in August 2018, and 784 at endline in January 2019, at 60 sites in the Western Highlands of the country. A small sub-sample of participants (the largest logistically feasible) was re-tested at endline within two weeks of full scale administration (57 youth in Uganda, 126 youth in Guatemala).

The **exploratory factor analysis (EFA)** performed on the initial round of baseline data from Uganda returned a 4-factor solution as the best fit, accounting for 87% of the variance. The 4-factor solution is similar to the original theoretical framework underlining the first draft of the tool which hypothesized five subscales constructed around five skills: positive self-concept, self-control, higher order thinking skills, social skills, and communication. In the 4-factor structure, Factor 1 reflects "Positive self-concept," Factor 2 reflect "Higher order thinking skills (HOTS)", Factor 3 "Negative self-concept", and Factor 4 "Social Skills and Communication." This scale structure was subsequently reexamined through a **confirmatory factor analysis (CFA),** using data from both Uganda endline, and Guatemala baseline and endline. The statistics employed to evaluate model fit were. The results of the CFA, including a range of statistics such as the Chi-square, Tucker-Lewis Index, comparative fit index, and the root mean square residual fit index, indicated good model fit in both countries.

An analysis of **internal consistency** (Cronbach's alphas) showed moderate to strong reliability in Uganda, values ranging from 0.55 to 0.79 at baseline and from 0.66 to 0.81 at endline. In Guatemala, Cronbach's alpha values were from 0.7 to 0.83 at baseline and from 0.7 to 0.86 at endline. Test-retest correlation coefficients ranged between .5 and .7 between Uganda and Guatemala.

Measurement Invariance (MI) Analysis

We examined MI on the following dimensions:

- Endline by country (Guatemala and Uganda)
- Guatemala by round (baseline and endline)
- Guatemala baseline by gender (girls and boys)
- Guatemala baseline by SES (low and high)
- Guatemala endline by gender
- Guatemala endline by SES
- Uganda endline by gender
- Uganda endline by SES

Our MI analysis found the tool to be invariant by all of these characteristics. All the configural-to-metric and metric-to-scalar changes in the model fit statistics are below these thresholds, which means the assessment demonstrates metric and scalar invariance for the constructs across all characteristics (see). **This indicates that the overall scale structure fits well in applications across contexts such as Uganda and Guatemala, and across the subgroups within these contexts.**

Differential Item Functioning

Similar in purpose to MI, differential item functioning (DIF) analysis considers how *items* (rather than scales and subscales) perform by different groups. If an item has differential functioning, it means that the item may be measuring different concepts depending on the group (rural or urban, for example) (Camilli and Shepard 1994). **Despite identifying DIF in some items, we did not observe a magnitude that would lead to noticeable differences at the construct level, reaffirming the stability of the tool in cross-cultural application**.

Change Over Time

While in neither Uganda nor Guatemala setting was an evaluation framework, a simple before-and-after measurement between points in time is the first step towards understanding whether the tool has the potential of capturing a program or time effect. For both samples, all the skills levels decreased, except for negative self-concept, which increased. The changes ranged from 0.04 to 0.13, although it is not significant for HOTS or social and communication skills for the panel sample. Because there is no comparison group, it is not possible to attribute this change to the program, however, the conclusion remains that the scales on positive and negative self-concept have registered change between the baseline and endline administrations.

Predictive Validity

We used the following four outcomes for our validation analysis: 1) "Employment Score", a variable indicating success in the workforce; 2) "Ever had sex", a variable indicating if the youth ever had sex; 3) "Any Disability", a variable indicating if the youth reported having any disability; and 4) "Any Violence", a variable indicating if the youth reported perpetrating any kind of violence (physical, verbal, or emotional). **Overall, although the magnitude of the correlation is modest for some of the outcomes, they are all in the expected direction.**

Anchoring Vignettes Adjustment

A set of anchoring vignettes (high and low on each scale) was included in the tool for each scale, for subsequent adjustment analysis. For AV adjustment, each item is rescored based on their position relative to the AV scoring. Overall, responses are in the direction we would expect, where respondents rated the High AV higher than the Low AV for more than 80% of the cases, with the exception of *Communication* for Uganda endline. This indicates that the level of engagement and comprehension is very high among respondents. Looking at correlations of the AV-adjusted scores with external variables, we find that the direction of the correlations is very similar to the ones obtained from the non-adjusted scores. However, the magnitude of the correlation is considerably lower, indicating the AV-adjusted scores are not out-performing the non-adjusted scores.

We conclude that it is not necessary to use AVs for cross-country comparisons between the two sites where the pilot was implemented. However, users may want to perform an AV analysis if they are conducting cross-country comparison between countries with contexts dissimilar to our two sites. We also recommend using the AVs to investigate engagement and comprehension of the survey.

Program Staff Assessment

We first "mapped" the items in the program staff tool to the youth tool factors based on the item content. After mapping each item to its corresponding youth item, we created three sub-scales for the program staff tool based on how the program staff items map onto each item within each youth factor. Next, we analyzed the correlations between the youth factors and each version of the program staff factor.

Correlations between the youth and program staff factors are generally low for both countries and at both times. The highest absolute value of a correlation is .15. Looking at changes in the correlation coefficients between baseline and endline, in Uganda, we see some increases and sign changes in the expected direction. In Guatemala, the changes from baseline to endline in the correlation coefficients present more of a mixed bag. Looking at the Cronbach's alphas from the Uganda data, reliabilities ranged widely, from .13 to .79 at baseline and from .24 to .75 at endline. For Guatemala, reliabilities also ranged widely, from .19 to .87 at baseline and .37 to .84 at endline.

Overall, we find that the results of the youth and program staff data together are difficult to interpret. On their own, however, some of the program staff tool subscales seem to hold together reliably. We would recommend further testing and revisions to this tool before using it in a program setting, given that it does not currently seem to add analytical value.

Analysis of Item Order

We tested whether the order of the items—and specifically whether certain items fall at the beginning or end of the assessment—affects how youth rate their skill levels by randomly assigning half of the sampled youth at each administration of the tool in both countries "Form A" and the other half "Form B." We find some evidence that the order of the items may affect youth responses, given that, in some cases, youth rate themselves differently depending on where an item is placed in the survey. However, we recommend further research to test this assumption.

Analysis of Enumerator Characteristics

Our analysis also considers the potential effects of enumerator gender, age, and research experience on youth soft skills and validation outcomes. Our findings suggest that enumerators' gender, age, and research experience all affect youth's responses, although it is unclear why.

Conclusion

In sum, the YouthPower Action Youth Soft Skills measurement tool presents a validated assessment of youth soft skills such as positive self-concept, negative self-concept, higher order thinking skills, and social and communication skills. The assessment can be used to predict youth outcomes in key areas and measure change in the level of soft skills over time, in as short as a few months. Further testing is necessary to determine validity in contexts substantially different than the youth programs in which the assessment was validated, with other youth outcomes, and in a causal inference framework.

Recommendations

For Tool Development

Our experience in developing this instrument reveals several key lessons.

1) **Contextualization is a critical first step**. The cognitive testing that we conducted in Uganda and Guatemala revealed critical information that informed our initial revisions and helped us to interpret several confounding findings.
2) **Less is more.** Overall, the tool performed better once we radically simplified the item wording and structure. We also found that additional tools (in our case, the program staff tool) may not provide additional analytical clarity.

For Implementation

We recommend that implementers seeking to measure program participants' soft skills focus on the following areas:

1) Implementers should build in sufficient time and resources in the program cycle to conduct baseline (and endline, where relevant) assessments—and to contextualize the assessment for the program and country context.
2) Implementers should also take a wide view of monitoring and evaluation processes **to ensure that participants do not experience survey fatigue**.

For Research

We recommend that future research studies focus on the following gaps:

1) Testing of this tool in the context of an aligned soft skills intervention, with a control group in order to more clearly isolate where change in skill levels may be coming from.
2) Further research on the pathways linking skills and outcomes.
3) **Testing and validation of this tool in additional contexts** in order to better understand cross-cultural differences in how skills change over time, how skills relate to key outcomes, and how skills relate to other key demographics.
4) Further research should be conducted on the effect of enumerator gender, age, and research experience in different contexts.

Introduction

Under USAID YouthPower Action, FHI 360 was asked to develop and validate an assessment of youth soft skills crucial to workforce success, violence prevention, and positive decision-making around reproductive health. Between 2017 and 2019, FHI 360 developed and pilot-tested an assessment instrument for administration with youth in lower resource environments that were the focus of USAID youth development programs. The design and testing included two instruments: a youth self-assessment, as well as a third-party assessment, intended for use with youth program mentors or facilitators.

The target group for the measurement tool is youth aged 15-19 years, both in and out-of-school, who are typically the beneficiaries of USAID-sponsored youth programs. End users of the tool were to include program staff working with youth beneficiaries and their caregivers, and donors seeking to assess youth's soft skills. The new soft skills assessment tool is intended for application in USAID-funded youth development programs that are required to assess youth soft skills levels.

The YouthPower Action Youth Skills Assessment is designed for one-on-one administration to youth by program monitoring and evaluation (M&E) staff in program and community settings. The duration of administration is approximately 45 minutes per respondent. While program staff may choose to collect data on electronic devices, paper-based administration and scoring is also possible. This report, written with M&E and research generalists in mind, discusses the process of the instrument development, the results of validation in Uganda and Guatemala, as well as the implications for further application and validation.

Background

Evidence across fields and disciplines highlights the importance of soft skills to long-term education, employment, health, and violence prevention outcomes (Deming 2015; Almlund et al., 2011; Heckman et al., 2006; Carneiro et al. 2007). At the same time, investments in school-based, out-of-school, and workplace-based programs and activities that promote developing soft skills among different groups of adolescents and young adults have grown significantly worldwide. USAID, through YouthPower Action, sought to expand the knowledge base on what soft skills were important for future youth success, and to encourage youth programs to adopt effective approaches to fostering them. At USAID's request, YouthPower Action, implemented by FHI 360, carried out a literature review to identify key soft skills for cross-sectoral youth outcomes, and subsequently embarked on a process for the development and validation of a new tool that would measure soft skills in a variety of cultural settings.

The literature review, covering 223 studies, sought to understand whether specific soft skills correlated with certain youth development outcomes—namely, workforce success, violence prevention, and reproductive health. Researchers and practitioners from multiple disciplines often referred to a general body of skills, whether using the term "soft skills", "socio-emotional skills", "life skills", or "21st century skills". The review sought to identify whether discrete, measurable skills, such as communication, critical thinking, and self-esteem, were predictive of positive youth outcomes. Based on the literature covered in the review, the research team proposed a set of skills – *positive self-concept, self-control, and higher order thinking skills* – that were found to correlate with the three outcomes of interest. The review (Gates et al., 2016; Lippman et al., 2015) recommended that youth development programs target these skills.

Following the literature review and agreement on the overall assessment framework covering the three key soft skills, YouthPower Action created an analysis of existing measurement assessments, to determine which, if any, could be adapted for use across USAID youth programs. The resulting repository of close to 300 instruments targeting youth between 12 and 29 years old. The team then reviewed each tool based on a set of criteria, including evidence of use by international youth

development programs; evidence of validity; relevant validation sample; evidence of use with youth development outcomes of interest (workforce, RH, and violence prevention outcomes); evidence of reliability; and evidence of international usage. The review revealed that no single instrument encompassed all key skills identified as most important across three domains (workforce, violence prevention, and RH); was suitable for measuring change in skill levels over time among youth in international development programs; and met the other key criteria for use by youth programs (including ease of administration, validity, and reliability). YouthPower Action sought to address this gap, through the development of a new instrument that builds on existing literature but meet the criteria for wide use across USAID youth programs.

Methodology

In addition to addressing gaps in the field of soft skills measurement, the tool also intends to measure at least the three key soft skills referred to above among youth aged 15-19 years, both in and out-of-school, who are typically the beneficiaries of USAID-sponsored youth programs. End users of the tool include program staff working with youth beneficiaries and their caregivers and donors seeking to assess youth's soft skills. The soft skills assessment tool will benefit USAID-funded youth development programs that are currently required to assess youth soft skills levels but have no appropriate measurement tool for doing so.

The key objective of the measurement study under YouthPower Action was to develop and validate a tool that would measure the three key soft skills seen to be correlating with positive youth outcomes – workforce success, violence prevention, and reproductive health – that would also be: a) targeted to youth ages 15-19; b) low cost and easy to administer, c) could measure change over time, and d) consistent enough to measure across cultural contexts.

The instrument development consisted of the following stages:

5. **Initial instrument design.** Building on the repository of existing tools, YouthPower Action identified items that scored the highest across a range of criteria, including whether the item was conceptualized as a measure of one of the three key soft skills, and whether it was previously in non-OECD settings. This process resulted in an initial version of the assessment, which was later rolled out for field validation.
6. **Field testing.** Two separate field validation processes tool place with youth programs in Uganda and Guatemala. Each field validation consisted of the following:
    a. **Cognitive testing.** Qualitative cognitive interviewing with prospective respondents sought to establish item comprehension and ensure that items and response options were appropriate for the context and that their phrasing clear to the target group of youth.
    b. **Field administration**. One-on-one surveys were administered to larger samples of youth in the two countries. The respondents were beneficiaries of youth programs. Testing was administered at two points in time, at the beginning and the end of a program intervention. At point 2 (endline), participants were split into a panel sample (same youth as at baseline) and new sample, to assess any effects of prior participation in the survey. At point 2, a subset of participants was also re-interviewed to assess the tool's test-retest reliability.
7. **Analysis and adaptation.** Two rounds of analysis and adaptation took place throughout the period of performance, one after the initial baseline in Uganda, followed by further extended analysis after a revised tool was administered in Guatemala and a repeat administration in Uganda.

a. **Initial revisions.** After the initial administration in Uganda, the team made revisions to the tool to further improve its clarity and minimize challenges with comprehension. Low-performing items were removed from the tool.
b. **Re-administration.** A revised version was piloted in Guatemala and re-administered in Uganda and Guatemala at the end of each intervention.

8. **Validation analysis.** Upon completion of the field stage, a series of analyses were performed to establish the following:
    a. **Scale structure.** Exploratory and confirmatory factor analyses were administered to identify the optimal scale structure for the instrument. While a theory-driven structure was assumed based on prior literature, items were found to load somewhat differently onto the framework than had been hypothesized. Adjustments reflected this empirically driven structure.
    b. **Reliability.** At the endline stage for each field site, the tests included a test-retest analysis with a limited sample to examine the stability of the scale measurements within a short period of time.
    c. **Predictive validity.** Each of the scales were examined for correlations with outcomes of interest, as well as key youth demographic characteristics, to examine whether the scales were predictive of outcomes as hypothesized and could be predicted by youth background.
    d. **Change over time analysis.** For the second field site in Guatemala, the team examined whether the tool was capturing change between the baseline and endline administrations.
    e. **Cross-cultural comparability**. A series of analyses examined whether the scales within the tool were stable enough across contexts to warrant cross-cultural application with limited adaptation. This involved measurement invariance analysis, differential item functioning analysis, and analysis of scale performance with anchoring vignettes (AV adjustment).

The study team intentionally prioritized the development, testing, and revision of the youth self-assessment over the third-party assessment, or program staff tool. The team made only minor revisions to the program staff assessment. Most of the revisions took place after the team analyzed data from the Uganda baseline administration of the assessment. At this point, the team removed any items from the program staff assessment that no longer had a corresponding item in the youth self-assessment. For more information on the development and analysis of the program staff tool, see the section that begins on page 22.

Below, the report provides a more detailed description of each of these steps and presents the results.

Instrument Design

Youth Assessment. Drawing from an item bank constructed from the repository of existing assessments, the initial version of the youth assessment included items that measured the thirteen sub-skills or sub-domains within three soft skills as described in the "Key Soft Skills" report. Items were subsequently modified to fit the developing country context, replacing references to items and situations that were not readily available to the majority of beneficiaries of USAID youth programs. Because the tool also had to be applicable to youth who were not in school, items that referenced school and issues related to being a student were also modified.

The response options for each item were initially designed as a 6-point behavioral frequency scale that asked youth to assess "how often" they act out a behavior that is associated with a particular skill (e.g., "How often do you think things through before you do them?"). Response options ranged from "Always

or Almost Always" to "Never or Almost Never". This was hypothesized as a way of obtaining greater precision and clarity in the responses. Subsequently, this decision was revised, following initial results from Uganda, and a simple 4-point endorsement scale, from "Strongly Disagree" to "Strongly Agree".

The initial version of the tool also included a group of items we refer to as "importance items", intended to assess how youth valued each of the measured skills, with the hypothesis that youth who value a certain skill were more likely to work to improve that skill. However, as the report details below, importance items were not found to be reliable or predictive of the actual scales and were later dropped.

To address the question of cross-cultural comparability, a series of *anchoring vignettes (AV's),* intended to circumvent cultural response bias, were included in the youth self-report component of the tool. AVs present hypothetical situations and people that illustrate skill levels, followed by a series of response options, one of which is correct. The respondent's rating of the AV is used to examine a consistent response bias, which, if captured can be addressed through an adjustment process. compared to the respondent's assessments of the hypothetical people described in the vignette(s). In the YouthPower Soft Skills assessment, one AV was drafted for each skill construct.

Finally, the youth assessment tool also included a module of socio-economic status proxies, a module on disability status (following the Washington Group Short Set of Questions), and a series of items measuring outcomes of interest: violence prevention, workforce success, and reproductive health. All items, including the outcomes, were designed as self-reported scales, with an ordinal scale measuring the frequency of the behavior, or the level to which the respondent agreed with a statement.

Program Staff Tool. In addition to the youth self-report instrument, YouthPower Action developed a third-party assessment tool in response to findings from the measurement tool review and feedback from a group of technical experts[2] that many existing soft skills tools suffer from self-report bias. The intention of the third-party assessment tool, which we envisioned being used with youth program mentors or facilitators, was to provide another, ideally more objective, measure of youth soft skills. To develop this instrument, the research team drafted items that corresponded to four out of five of the skills constructs—self-control, higher order thinking skills, social skills, and communication. Positive self-concept was not considered to be appropriate for measurement through third-party assessment.

Field Testing

As noted above, the tool was tested in two different sites: in Uganda with the Educate! Program, a non-US Government program for youth in secondary schools, and in Guatemala with the Proyecto Puentes, a USAID funded program implemented by World Vision in the Western Highlands of Guatemala that delivers life skills and technical and vocational training among 15-24-year-olds, both in-school and out-of-school. The data collection phase consisted of two primary research activities: 1) qualitative cognitive interviews to assess the construct validity of the items and response options; and 2) baseline data collection with a large sample of youth to assess the performance of the tool. Prior to the cognitive interview phase, we asked the in-country researcher and the Educate! monitoring and evaluation team to review the tool and response options for cultural appropriateness. They provided several minor revisions to the item wording.

[2] **Technical Advisory Group**, working with the team throughout 2017 on instrument development. See Acknowledgements for group members.

Site 1: Educate! Program in Uganda

Educate! is an after-school soft skills non-USG development program providing entrepreneurship and life skills training for secondary school-aged youth across Uganda. Operating in Uganda since 2010, Educate! gradually expanded to nearly 600 schools, with a model that integrates training as after-school activities for a select group of eligible youth, or scholars. Educate! scholars were selected through an admissions process, where youth had to demonstrate motivation and entrepreneurship and leadership potential. In addition to the structured after-school sessions, Educate! students also formed business clubs, with the support of their mentors, where they practiced the skills learned in the classroom. With these selection criteria in place, Educate! Scholars differed somewhat from the eventual target population for the soft skills measurement tool, which was intended for disadvantaged youth. However, the breadth of the program, its scale in Uganda, as well as its implementation calendar, made it an applicable setting in which to test the new instrument.

Site 2: Proyecto Puentes in Guatemala

Proyecto Puentes is a youth development program implemented by World Vision in the Western Highlands of Guatemala that delivers life skills and technical and vocational training among 15-24-year-olds, both in-school and out-of-school. The target group for Puentes is disadvantaged, mostly indigenous youth that seek to improve their skills for subsequent application in the job market. Unlike Educate! scholars, Puentes's target beneficiaries were not enrolled in secondary schools or other education institutions, and often worked a variety of jobs, as well as helped take care of other family members. This made youth recruitment and follow-up more challenging for enumerators, who had to schedule specific times for youth to appear for interviews.

Cognitive Testing

Cognitive testing took place in both locations. In Uganda, this included a series of in-depth qualitative cognitive interviews with 50 youth, a pretest of the youth tool with 23 youth, and a pretest of the program staff tool with five program staff at six schools (three urban and three rural) in three locations in Uganda—Kampala, Jinja, and Gulu—during October-November 2017. The qualitative cognitive interviews[3] served to assess the content validity of the items, various response options, and the overall structure of the tool; the full test-run of the tool to ensure smooth processes for electronic data capture and analysis. This process yielded useful lessons about the tool, relating to students' preferences on the response options and students' understanding of the items—in particular, how cultural norms and values affected their interpretation of the items. Through this learning, the team settled on a set of response options, revised item wording, revised the content of some questions (for example, by adding more specific examples), and removed some questions altogether.

In Guatemala, the assessment instruments (both the youth self-assessment and the third-party instrument) were administered in Spanish. A translation from English to Spanish, and backtranslation were administered prior to cognitive testing, to generate a working version. The formative research component followed the process used in Uganda, consisting of in-depth qualitative cognitive interviews in Spanish that assessed the content validity of the items and the overall structure of the tool. In August 2018, three data collectors interviewed[4] approximately 55 youth in three communities in the Western

[3] The five-member study team consisted of three Ugandans and two Americans. Most of the interviews were conducted in English. In some cases, in order to clarify difficult concepts, the Ugandan enumerators mixed English and local languages. At one school, the three Ugandan enumerators had to rely heavily on local languages due to the students' lower-than-average English language levels.

[4] All three data collectors were fluent in Spanish and conducted the interviews in Spanish.

Highlands in Guatemala. Through a rapid analysis of the cognitive interview data, the team identified problematic words and phrasing and revised the Spanish version of the tool accordingly.

In Uganda, FHI 360 contracted the Centre for Social Research (CRS) to carry out the data collection for both baseline and endline. Prior to the baseline data collection in March 2018, a two-person team from FHI 360 conducted a five-day training for the enumerators on tool development, study design, and recruitment and sampling; use of the Samsung tablets and the software program used for data collection the tools; and research ethics. For sampling, FHI 360 obtained from Educate! a list of participating secondary schools in their program and randomly selected 59 schools from that list to participate in the data collection. At each school, 19 youth, including both Educate! scholars and age-eligible non-scholars were sampled for the interviews.

In Guatemala, a Honduras-based firm Espiralica was contracted to conduct baseline data collection at 60 program sites in September 2018. Prior to baseline data collection in September 2018, a two-person team from FHI 360 conducted a five-day training for the enumerators that covered the same information as that covered in the Uganda training. For sampling, the program sites were selected randomly from a list of program sites provided by the Proyecto Puentes team. Local program coordinators for Proyecto Puentes then randomly selected youth from their attendance lists who met the study inclusion criteria (being a participant in Proyecto Puentes and between 15 and 19 years old) and asked them to come to the youth center on a certain day and time.

Endline Data Collection

Refresher trainings were conducted for both Uganda and Guatemala teams prior to the endline data collection. In Uganda, CSR carried out endline data in November 2018 at the same 59 schools, with a split sample of youth that had been tested at baseline, and new participants, to test any prior participation effects. For the endline data collection in Guatemala, YouthPower Action procured the services of an enumeration firm Khanti, based in Guatemala. The Khanti research teams of conducted endline data collection in January-February 2019 at the same 60 sites.

Further, for test-retest analysis, CSR teams re-administered the tool with a subsample of 57 youth, who had already been interviewed once at endline, to capture the tool's test-retest reliability. In Guatemala, 126 youth were re-interviewed for test-retest analysis.

Data Analysis

Scale Construction

The team pursued an iterative process in scale construction, starting with an initial instrument drafted and cognitively tested in each country, and subsequent revisions to reduce respondent burden, strengthen clarity, and minimize measurement error. The initial batch of data from the Uganda baseline data pointed to issues with comprehension and interpretability of some of the items, as manifest in low item-rest[5] and item-test correlations, and low sub-scale Cronbach's alphas.

The following instrument revisions were undertaken between the Uganda baseline and the Guatemala baseline administrations:

[5] Item-rest correlation, or item-total correlation, refers to the correlation between the item and the sum of the rest of the item scores.

1. Item stem revisions aimed at streamlining the language. For example:
    a. How often do you believe that: you are good at learning something new? → I'm good at learning new things.
    b. How often do you do something risky because of peer pressure? → If my friends are doing something risky, I will do it with them.
    c. (In response to a prompt): How often did you take action to solve the problem → I took action to solve the problem.
2. Response options revisions from behavioral frequency formulation to endorsements. For example, "How often do you believe that: you are good at learning something new?" [Response options: Almost Never, Rarely, Sometimes, Often, Almost Always] was revised to "I'm good at learning new things" [Strongly agree; agree; disagree; strongly disagree].
3. "Importance" items (a set of items asking students how important they thought a skill was), were removed for their lack of analytical value. Most of the responses to these questions clustered at the extreme ends (i.e., nearly all youth responded that they the skill was either "important" or "very important").
4. Removal of low-performing items. Items that loaded low on all of the factors in our exploratory factor analysis or did not correlate with other items in their scale – indicating challenges with comprehension or lack of relevance for participants – were removed. These items were typically wordy or indirectly phrased.

The revised version was subsequently applied in the baseline testing in Guatemala, and the endline administrations in both Uganda and Guatemala. The following describes the scale structure, results of the statistical reliability and validity testing, and emerging insights on the predictive validity of the tool.

Exploratory Factor Analysis (EFA)

An exploratory factor analysis (EFA), which identifies the most optimal scale composition for a survey or assessment, was a critical component of our study. **The EFA which we ran returned a 4-factor solution as the best fit for the data, accounting for 87% of the variance.** The factor analysis summary is shown in Table 1. Factor loadings and uniqueness are shown in the appendix. Overall, items presenting high loadings and cross-loading occurred for only one item (problem-solving 4). Based on the factor analysis results, 13 items that did not load onto any of the scales were excluded from the tool, as noted above.

The 4-factor solution is similar to the original theoretical framework. Factor 1 reflects "Positive self-concept," Factor 2 reflect "Higher order thinking skills (HOTS)", Factor 3 "Negative self-concept", and Factor 4 "Social Skills and Communication." The items that loaded onto factor 1 fall under our originally hypothesized construct "positive self-concept." We've summarized how the items from the 5-factor structure map onto each of the 4 factors in Table A. 1.

The only notable departure from the originally hypothesized structure is the emergence of Factor 3, which reflects the negative self-concept as distinct from the positive self-concept continuum measured by Factor 1. This distinction between negative self-concept and positive self-concept is supported by evidence from the psychology literature, which consistently finds two separate personality dimensions in negative (distress and general negativity, including anger, disgust, fear, and nervousness) and positive affect (enthusiasm, activeness, alertness) (Watson, Clark, and Tellegen, 1988; Watson and Tellegen, 1985). Similarly, Tellegen and Waller's (1987) analysis finds seven major personality dimensions, which include two dimensions that reflect how one feels about oneself: positive valence (for example, "excellent" vs. "ordinary") and negative valence (for example, "evil" vs "decent.")

The items that loaded onto Factor 4 include items from our original conceptualizations of social skills and communication and items that measure how well youth ask for help when they try to solve problems. Examples include: "I get along well with people from different backgrounds" and "I write well." Our initial distinction of communication skills from social skills came from the way communication skills was defined in the workforce literature as a separate skill (Lippman et al., 2015). However, it is not surprising that these two factors converged, given the overlap in the behaviors associated with these skills, such as participating in a team, asserting oneself appropriately to resolve a conflict, and complimenting others.[6]

Table 1. Eigenvalues, Percentages of Variance, and Cumulative Percentages – Uganda and Guatemala Baseline

		Uganda		
		Eigenvalue	% Variance	Cumulative %
Factor1	Positive self-concept	7.32	0.59	0.59
Factor2	HOTS	1.63	0.13	0.72
Factor3	Negative self-concept	1.15	0.09	0.81
Factor4	Social & communication skills	0.67	0.05	0.87
		Guatemala		
		Eigenvalue	% Variance	Cumulative %
Factor1	Positive self-concept	7.30	0.55	0.55
Factor2	Negative self-concept	1.67	0.13	0.68
Factor3	HOTS	1.43	0.11	0.79
Factor4	Social & communication skills	0.90	0.07	0.86

The revised 50 items of the Guatemala baseline instrument were subjected to a second exploratory factor analysis. The Kaiser's criterion of eigenvalues greater than 1 and the parallel analysis yielded a four-factor solution as the best fit for the data, accounting for 86% of the variance. The four factors presented a similar structed from the EFA performed in Uganda baseline, except that negative self-concept accounted for more of the variance than HOTS. Factor loadings and uniqueness are shown in the appendix.

Model Fit Assessment through Confirmatory Factor Analysis (CFA)

Following the scale structure that emerged through the exploratory factor analysis, the team reexamined the fit of the new scales through a confirmatory factor analysis (CFA). **The results of the CFA confirmed a good model fit in both countries**. The team analyzed model fit through maximum likelihood estimation in the R package *lavaan* (Rosseel, 2012), with ordinal response scales. The statistics employed to evaluate model fit were Chi-square, Tucker-Lewis Index (TLI), comparative fit index (CFI), and root mean square residual fit index (RMSEA).[7]

Because the Chi-square is sensitive to large sample sizes and may reject well-fitting models, our model fit assessment placed more emphasis on the other statistics. According to Hu and Bentler (1999), CFI and TLI statistics greater than 0.9 are considered to be an "adequate" model fit, whereas values greater

[6] See Harvard's Explore SEL Taxonomy for more examples.
[7] Descriptions of these model fit statistics can be obtained from Bollen (1989), Hoyle (1995), and Hu and Bentler (1999).

than 0.95 are considered as a "good" model fit; fit indices for RMSEA less than 0.8 are considered "good". The comparative fit index (0.98 for Uganda and 0.96 for Guatemala); the Tucker-Lewis fit index (0.98 for Uganda and 0.96 for Guatemala); and the root mean square residual fit index (0.03 for Uganda and 0.05 for Guatemala) indicate a good fit between the model and the observed data, as shown in Table 2. No post-hoc modifications were indicated from the analysis because of the goodness-of-fit indexes, and the residual analysis did not indicate any problems.

Table 2. CFA Goodness-of-fit indicators of models for Uganda and Guatemala

	Uganda	Guatemala
chi2	2119	3150.4
df	1074	1074
p > chi2	0	0
CFI	0.98	0.96
TLI	0.98	0.96
RMSE	0.03	0.05
90% CI, lower bound	0.03	0.05
upper bound	0.03	0.05
Observations	**1010**	**784**

Scale Reliability

One of the key desired characteristics of the resulting Soft Skills measurement tool is its internal consistency, its test-retest reliability, and its invariance to different population subgroups and different cross-cultural contexts. This section presents the analysis of internal consistency, test-rest reliability, and cross-group and cross-cultural measurement invariance analysis and differential item functioning analysis. **These analyses point to the general conclusion of the instrument holding up, both in reliability and consistency, and comparability within and across the Uganda and Guatemala samples**.

Internal consistency

The Cronbach's alpha statistic is a common initial measure of scale internal consistency, or the strength of the relationships between items within a scale. Table 3 shows Cronbach's alpha statistics for the 4-scales obtained from Guatemala's baseline EFA analysis for both countries at baseline and endline. For Uganda, values ranged from 0.55 to 0.79 at baseline and from 0.66 to 0.81 at endline; and for Guatemala, Cronbach's alpha values were from 0.7 to 0.83 at baseline and from 0.7 to 0.86 at endline, indicating a substantial level of internal consistency. The increase in Cronbach's alphas for Uganda between baseline and endline reflects the improvement in the tool after the item stem and response scale revisions.

Table 3. Cronbach's alpha reliability for Uganda and Guatemala, baseline and endline

Dimension	Uganda		Guatemala	
	Baseline*	Endline	Baseline	Endline
Positive self-concept	0.79	0.81	0.83	0.86
Negative self-concept	0.62	0.66	0.7	0.76
HOTS	0.72	0.69	0.7	0.7
Social & communication skills	0.55	0.66	0.71	0.72
Sample Size	**1098**	**1010**	**794**	**784**

*Note: scores for Uganda baseline were calculated using a five-point scale, while scores for Uganda endline and Guatemala were calculated using a four-point scale.

Note that these initial measures are based on raw scale scores. An Anchoring Vignettes adjustment process, which may be applied as part of instrument scoring, is described below on p. 23, including its effect on the Cronbach's alpha statistics.

Test-Retest Reliability

Test-retest reliability analysis is an assessment of measurement stability over a short period of time. In this study, repeat administration was performed for a subsample of 57 youth at Uganda endline and 126 youth at Guatemala endline over an interval of two weeks, with the same enumerators. Table 4 shows that the test-retest correlation coefficients ranged between .5 and .7 between Uganda and Guatemala.

Table 4. Test-retest reliability coefficients

Dimension	Uganda	Guatemala
Positive self-concept	0.7	0.7
Negative self-concept	0.58	0.65
HOTS	0.65	0.64
Social & communication skills	0.56	0.7
Sample Size	**57**	**126**

The test-rest statistics displayed at the scale level, while below the conventionally desired 0.75 level, are comparable to those reported on other soft skills measurement tools, such as the General Self-Efficacy Scale (.45 - .75; (Schwarzer & Jerusalem 1995); the Responses to Stress Questionnaire (.49 to .76 for the 19 parcels and .69 to .81 for the five factors (Conner-Smith et al. 2000); and the Strengths and Difficulties Questionnaire: Parent or Teacher version (for the parent version: .54 - .81; for the teacher version .58 - .8 (see Stone et al. 2015)). The knowledge base on reasonable test-retest reliability statistics to expect on soft skills assessment is still limited; however, it appears that the levels reported on the YouthPower tool follow the trend of these prior assessments.

Measurement Invariance Analysis

Measurement invariance (MI) analysis is an even further step in reliability testing. Measurement invariance explores the extent to which the soft skills constructs, as they emerged in the 4-factor scale structure described above, have the same meaning for different groups of participants. We examined MI on the following dimensions:

- Endline by country (Guatemala and Uganda)
- Guatemala by round (baseline and endline)
- Guatemala baseline by gender (girls and boys)
- Guatemala baseline by SES (low and high)
- Guatemala endline by gender
- Guatemala endline by SES
- Uganda endline by gender
- Uganda endline by SES

For example, using the endline data, the MI analysis by country helps us understand whether youth from these two different contexts have different understandings of the assessment items; the MI analysis by

gender allows us to understand whether young men and young women have different understandings of the assessment items; and so on. Our MI analysis found the tool to be invariant by all of these characteristics. All the configural-to-metric and metric-to-scalar changes in the model fit statistics are below these thresholds, which means the assessment demonstrates metric and scalar invariance for the constructs across all characteristics (Appendix 4). **This indicates that the overall scale structure fits well in applications across contexts such as Uganda and Guatemala, and across the subgroups within these contexts.** For more information on the MI analysis process, see Appendix 3.

Differential Item Functioning

Similar in purpose to MI, differential item functioning (DIF) analysis considers how *items* (rather than scales and subscales) perform by different groups. If an item has differential functioning, it means that the item may be measuring different concepts depending on the group (rural or urban, for example) (Camilli and Shepard 1994).

Despite identifying DIF in some items, we did not observe a magnitude that would lead to noticeable differences at the construct level, reaffirming the stability of the tool in cross-cultural application. The results are summarized in Appendix 4. Each cell contains the number of items flagged based on (1) a likelihood ratio (LR) Chi-square test, (2) an R squared test, and (3) Beta test. In the by-country comparison using the endline data, only two items were flagged by all three criteria, item "self_esteem2" in factor 1 and item "critical_thinking2" in factor 3 (see Appendix 5). The items flagged for DIF in the comparisons by gender and SES are listed in in the Appendix. In each case, a comparison of the trait estimates for the factors involved revealed broad overlap in the distributions compared, indicating that generally, items performed similarly for different subgroups.

Change Over Time

A crucial element of the tool validation is the assessment of its ability to capture change in soft skills over time, plausibly affected by an intervention. While in neither Uganda nor Guatemala setting was an evaluation framework, a simple before-and-after measurement between points in time is the first step towards understanding whether the tool has the potential of capturing a program or time effect. This section presents descriptive statistics on the scale for Uganda and Guatemala, at baseline and endline, as well as average frequency of response option by scale, context and time-period. First, we present descriptive statistics by groups, and validation with external variables is presented next, followed by the analysis of change over time, in the Guatemala context.

Table 5. Mean differences and change over time for Guatemala baseline and endline

	Guatemala					
	All Sample			Panel Sample		
	Baseline	Endline	Diff (E-B)	Baseline	Endline	Diff (E-B)
Positive Self Concept	3.38	3.27	-0.11*	3.38	3.28	-0.10*
Negative Self Concept	2.04	2.17	0.13*	2.04	2.15	0.11*
HOTS	2.98	2.94	-0.04*	2.99	2.96	-0.04
Social & Communication skills	2.97	2.93	-0.04*	2.98	2.95	-0.03
Observations	**794**	**784**		**450**	**450**	

Table 5 shows how the scales changed over time in Guatemala.[8] The first panel shows the full sample of students who were interview either at baseline or endline, while the second panel shows the sample of students who were surveyed both at baseline and endline (panel sample). For both samples, all the skills levels decreased, except for negative self-concept, which increased. The changes ranged from 0.04 to 0.13, although it is not significant for HOTS or social and communication skills for the panel sample. Because there is no comparison group, it is not possible to attribute this change to the program, however, the conclusion remains that the scales on positive and negative self-concept have registered change between the baseline and endline administrations.

Predictive Validity

The tool includes data on youth characteristics and potential behavioral outcomes, as a means of validation for the scales. The four validation outcomes are: 1) "Employment Score", a variable indicating success in the workforce; 2) "Ever had sex", a variable indicating if the youth ever had sex; 3) "Any Disability", a variable indicating if the youth reported having any disability; and 4) "Any Violence", a variable indicating if the youth reported perpetrating any kind of violence (physical, verbal, or emotional).

Table 6. Outcome Variable Definitions

Variable	Composition
Employment Score	Average of the following binary variables, coded 1 for "Yes" and 0 for "No": 1) In the last 3 months, did you receive payment for any work that you did? 2) Have you ever been interviewed for a job? 3) Have you ever received a job offer? 4) Are you running my own business?
Disability	Following the Washington Group Short Set of Questions, the list of disability questions included: difficulty seeing, even if wearing glasses; difficulty hearing, even if using a hearing aid; difficult walking or climbing steps; difficulty remembering or concentrating; difficulty (with self-care such as) bathing or dressing; and difficulty communicating, for example understanding or being understood. Response options included no difficulty; some difficulty; a lot of difficulty; cannot do at all. "Any Disability" variable assumes value 1 if students answered having at least some difficult to any of the question s and zero if they answered no difficulty to all of the questions.
Sex	Indicates whether the youth reported having had sexual activity.
Violence	Students were asked how many times in the past month they engaged in the following actions: insulted someone else's family (i.e. said something bad about them); made fun of or mocked someone else to make them angry; shamed or embarrassed someone to their face; not let someone be a part of your group anymore because you were upset or angry at them; and said mean things about someone to make others laugh. "Any Violence" variable assumes value 1 if students answered once or more for any of the questions, and zero if they answered zero for all of the questions.

Figure A. 1 – Figure A.24 show correlations between the scales and the validation outcomes. **Overall, although the magnitude of the correlation is modest for some of the outcomes, they are all in the expected direction.** Youth with higher positive self-concept, HOTS, and social and communication skills are more likely to have a higher employment score, while youth with a higher negative self-concept are less likely to have a higher employment score (although the latter relationship

[8] Because the tool changed from baseline to endline in Uganda, it is only possible to look at change over time using Guatemala data.

is not significant in Uganda). In Uganda, youth with higher positive self-concept are less likely to have ever had sex, while youth with a higher negative self-concept are more likely to have ever had sex. In Guatemala, youth with higher social and communication skills are less likely to have ever had sex. Looking at disability status, youth who report having "any disability" also report lower positive skills (positive self-concept, HOTS, and communication and social skills) and lower negative self-concept. Finally, youth who report having perpetrated "any violence" also report higher negative self-concept and lower positive skills (although the relationship is not significant for positive self-concept in Guatemala).

Table 7. Correlations between soft skills scales and external variables, Uganda (endline) and Guatemala (baseline)

	Positive Self Concept		**Negative Self Concept**		**HOTS**		**Communication & Social Skills**	
	Uganda	Guatemala	Uganda	Guatemala	Uganda	Guatemala	Uganda	Guatemala
Employment score	0.07** [0.03]	0.07** [0.03]	-0.04 [0.03]	-0.05* [0.03]	0.08*** [0.03]	0.05** [0.02]	0.06** [0.03]	0.05** [0.02]
Ever had sex	-0.08* [0.05]	0.04 [0.04]	0.09** [0.04]	-0.06 [0.04]	-0.05 [0.04]	0.01 [0.04]	-0.02 [0.04]	0.09*** [0.03]
Any Disability	-0.12** [0.05]	-0.18*** [0.06]	0.16*** [0.04]	0.16*** [0.06]	-0.13*** [0.05]	-0.24*** [0.05]	-0.13*** [0.04]	-0.21*** [0.05]
Any violence	-0.19*** [0.05]	-0.09 [0.06]	0.26*** [0.04]	0.15*** [0.06]	-0.18*** [0.04]	-0.17*** [0.05]	-0.26*** [0.04]	-0.17*** [0.05]
Observations	**1010**	**794**	**1010**	**794**	**1010**	**794**	**1010**	**794**

Notes: Significance is denoted as: * p < 0.1 ** p < 0.05 *** p < 0.01. Correlations at baseline shown for Guatemala and at endline for Uganda.

In sum, the scales appear to be predictive of the youth behaviors important for the validation, and these relationships are similar in magnitude across the two countries. Additional data, correlations with variables that measure other health-related behaviors, and visuals on the analysis and the correlations between the soft skills scales and outcomes of interest are provided in Appendix 1.

Anchoring Vignettes Adjustment

The anchoring vignettes (AV) technique is used to improve comparability among assessments of attitudes and preferences in self-report questionnaires (King & Wand, 2017). The method aims to assess response styles using short descriptions of hypothetical persons (vignettes) that vary systematically in the latent traits represented in the inventory. Respondents are requested to rate the persons described in the vignettes on an item similar to those used for the respondents' self-descriptions, adopting the same response format and rating scale (Primi et. al 2016). A secondary use of anchoring vignettes is to assess participants' comprehension of and engagement with the survey—if participants rate the vignettes correctly, this indicates good comprehension and engagement.

The application of anchoring vignettes usually involves attributing at least one vignette per item, if not more. Previous applications have used as many as 12 vignettes per self-assessment questions (King & Wang 2017). However, the addition of one or more questions for each self-assessment may increase the application costs considerably. Moreover, it can substantially increase the total time of the survey, adding to respondent burden. For these reasons, we opted to match two sets of AVs to many items, which may have impacted the validity of the AV-adjusted scores.

A set of anchoring vignettes (high and low on each scale) was included in the tool for each scale, for subsequent adjustment analysis. For AV adjustment, each item is rescored based on their position relative to the AV scoring: higher than High AV (score=5), same as High AV (score=4), between High

and Low AV (score=3), same as Low AV (score=2), and lower than Low AV (score=1).[9] The scale for the raw scores is 1 to 4; the scale for the AV adjusted scores is 1 to 5. This allows for a non-parametric adjustment that is not based on assumptions of an underlying distribution for each item. There are two independent AV adjustments for the Social and Communication Skills scale, due to having two sets of AVs for that scale. Appendix 6 includes more information on the scale means for the raw and anchoring vignette adjusted scores.

Figure A. 25 in the appendix shows the response patterns for each set of AVs. The first row shows the frequencies for Uganda endline, and the second and third rows show frequencies for Guatemala baseline and endline, respectively. The first bar, labeled *"1,2"* shows the frequency of responses when the respondents ordered the High AV higher than the Low AV; the second bar, labeled *"{1,2}"*, shows when responses were tied, which means respondents gave the same rating for the Low and the High AV; and the third bar shows the cases where the Low AV was rated higher than the High AV. Overall, responses are in the direction we would expect, where respondents rated the High AV higher than the Low AV for more than 80% of the cases, with the exception of *Communication* for Uganda endline. We can take this as an indication that the level of engagement and comprehension is very high among respondents, and the AVs are working as a good diagnostic that respondents are taking the survey seriously.

Table 8 shows Cronbach's alpha statistics for the 4-scales obtained from the AV-adjusted scores. Internal consistency values are stronger for AV-adjusted scores. However, Cronbach's alpha is not an adequate measure to evaluate AVs reliability, as pointed out by Daiver et al (2017). According to the authors, the AV approach relies on the assumptions that the vignettes are supposed to be invariant across respondents and the response to vignette prompts are supposed to be without error and strictly ordered. They show these assumptions are not always met and that higher Cronbach's alpha are obtained regardless of whether the assumptions are met or not.

Table 8. Cronbach's alpha AV reliability using AV-adjusted scores

Dimension	Uganda	Guatemala	
	Endline	Baseline	Endline
Positive self-concept	0.94	0.94	0.95
Negative self-concept	0.85	0.81	0.89
HOTS	0.89	0.86	0.92
Social & communication skills I	0.84	0.79	0.85
Social & communication skills II	0.86	0.81	0.87
Sample Size	**1098**	**794**	**784**

Correlating the AV adjusted score with outcome variables is a more appropriate method of exploring the validity of the vignettes (He et al. 2017; Kyllonen & Bertling 2014; Primi et al. 2016).

Table 9 shows correlations of the AV-adjusted scores with external variables. Although the direction of the correlations is very similar to the ones obtained from the non-adjusted scores, the magnitude of the

[9] The R package *anchors* was used to perform the AV adjustment (King & Lau 2008).

correlation is considerably lower, indicating the AV-adjusted scores are not out-performing the non-adjusted scores.

Table 9. Correlations between soft skills scales and external variables using AV-adjusted scores, Uganda and Guatemala

	Positive Self Concept		Negative Self Concept		HOTS		Communication & Social Skills I		Communication & Social Skills II	
	Uganda	Guatemala	Uganda	Guatemala	Uganda	Guatemala	Uganda	Guatemala	Uganda	Guatemala
Employment score	-0.00 [0.02]	0.01 [0.01]	-0.01 [0.02]	-0.00 [0.02]	0.02 [0.02]	-0.00 [0.01]	0.00 [0.01]	0.00 [0.01]	-0.02 [0.02]	-0.01 [0.01]
Ever had sex	-0.03 [0.02]	0.03 [0.02]	0.02 [0.02]	0.04* [0.02]	-0.00 [0.02]	-0.04** [0.02]	-0.01 [0.02]	-0.01 [0.02]	-0.00 [0.02]	0.03 [0.02]
Any Disability	-0.03 [0.03]	-0.06* [0.03]	0.03 [0.03]	0.05 [0.04]	-0.05* [0.03]	-0.03 [0.03]	-0.09*** [0.02]	-0.10*** [0.03]	-0.04* [0.03]	-0.10*** [0.03]
Any violence	-0.02 [0.03]	-0.06* [0.03]	0.07** [0.03]	0.05 [0.04]	-0.04 [0.03]	-0.03 [0.03]	-0.10*** [0.02]	-0.05* [0.03]	-0.11*** [0.02]	-0.11*** [0.03]
Observations	**1010**	**794**	**1010**	**794**	**1010**	**794**	**1010**	**794**	**1010**	**794**

Notes: Significance is denoted as: * p < 0.1 ** p < 0.05 *** p < 0.01. Correlations at baseline shown for Guatemala and at endline for Uganda.

Because the AV-adjusted scores are not bringing any improvement to the scales, we recommend the use of the non-adjusted scores for the main analysis. We do not find support for the need to use AVs for cross-country comparisons between the two sites where the pilot was implemented. However, users may want to perform an AV analysis if they are conducting cross-country comparison between countries with contexts dissimilar to our two sites. We also recommend using the AVs to investigate engagement and comprehension of the survey, as shown by Figure A. 25.

Program Staff Assessment

The first step to prepare for analysis of the program staff tool data was to "map" the items in the program staff tool to the youth tool factors based on the item content. The purpose of this mapping was to create subscales from the program staff tool that we could correlate with the youth tool. However, this was complicated by a few issues. First, some of the program staff items correspond to more than one of the four factors from the youth tool. For example, the item "How often does the youth think through things before doing them?" from the program staff tool intends to measure impulse control.
In the youth tool, there are two items that intend to measure impulse control: 1) I do things without thinking about them (impulses1) and 2) I think carefully before doing anything (impulses2). Thus, the program staff item on impulses maps onto two youth items.

Conversely, some items from the program staff tool have no equivalent in the youth tool. This is the case for the staff item on delayed gratification. While there were items on delayed gratification in the initial version of the youth tool (implemented at baseline in Uganda), these were deleted for subsequent iterations. There are also items that had no equivalent in either version of the youth tool (for example, "How often does the youth speak articulately?"). In both cases, these items were excluded from the analysis (see Appendix 7). for the mapping of the program staff tool items to the youth items and factors.)

After mapping each item to its corresponding youth item, we created three sub-scales for the program staff tool based on how the program staff items map onto each item within each youth factor.[10] (We refer to these at Subscales 2, 3, and 4, so that they correspond to the relevant youth factors (2, 3, and 4). Because the program staff tool items can map onto more than one youth item and thus youth factor (see Appendix 7), for our analysis, we created four different versions of the program staff subscales to capture all of the different possible groupings of the program staff items.

Next, we analyzed the correlations between the youth factors and each version of the program staff factor. We also generated Cronbach's alphas for each of the program staff factors. The results for each country at baseline and endline are displayed in the correlation matrices through Table A. 12 - A.15.

Correlations between the youth and program staff factors are generally low for both countries and at both times. The highest absolute value of a correlation is .15. The correlation coefficient of .15 (which is negative) is between Factor 4 of the youth tool and Subscale 4 of the staff tool and is not in the expected direction. The highest absolute value of a correlation that is in the expected direction is .12. This coefficient is negative, but this is the expected direction because it is between Youth Factor 3 (which contains all negatively worded items) and Program Staff Subscale 3.

Looking at changes in the correlation coefficients between baseline and endline, in Uganda, we see some increases and sign changes in the expected direction. The correlations between Youth Factor 2 and Program Staff Subscale 2 change from all negative to all positive from Time 1 to Time 2, indicating possibly slightly more relatedness between these questions at endline. We also see an increase in the (negative) correlations between Youth Factor 3 and Program Staff Subscale 3. The correlations between Factor 4 and Subscale 4 are more confounding – they are more negative at baseline and less negative at endline, but still negative.

In Guatemala, the changes from baseline to endline in the correlation coefficients present more of a mixed bag. The correlations between Youth Factor 2 and Program Staff Subscale 2 change from more negative to slightly less negative. Looking at Factor 3 and Subscale 3, all of the correlations are positive (which is unexpected)—some decrease and some increase from baseline to endline. All of the correlations between Factor 4 and Subscale 4 decrease from baseline to endline.

Looking at the Cronbach's alphas from the Uganda data, reliabilities ranged widely, from .13 to .79 at baseline and from .24 to .75 at endline. For Guatemala, reliabilities also ranged widely, from .19 to .87 at baseline and .37 to .84 at endline. Most reliabilities are above .5. For the Guatemala baseline data, all program staff subscales demonstrated reliabilities higher than .7, except for all versions of subscale 3. For the Guatemala endline data, all program staff subscales demonstrated reliabilities higher than .68, except for all versions of subscale 3. In Uganda, this pattern is only present at endline, where all subscales, except subscale 3, have reliabilities of at least .7. At baseline, the reliabilities are more mixed.

Overall, we find that the results of the youth and program staff data together are difficult to interpret. On their own, however, some of the program staff tool subscales seem to hold together reliably. We would recommend further testing and revisions to this tool before using it in a program setting, given that it does not currently seem to add analytical value.

---

[10] As we mention in the section on Tool Design, we did not develop Program Staff Tool items for positive self-concept, which is why there are only three sub-scales for the program staff tool (negative self-concept—which contains many of the items from the original theoretical "self-control" factor, HOTS, and social and communication skills).

Analysis of Item Order

We also tested whether the order of the items—and specifically whether certain items fall at the beginning or end of the assessment—affects how youth rate their skill levels. We tested this by randomly assigning half of the sampled youth at each administration of the tool in both countries "Form A" and the other half "Form B." In both form versions, the items are ordered by factors according to the original theoretical five-factor structure. In Form A, the factor order is positive self-concept, self-control, higher order thinking skills, communication, and social skills. In Form B, the factor order is reversed: social skills, communication, higher order thinking skills, self-control, and positive self-concept.

In Table 10 below, we present the results of t-tests that compare youth's mean scores for each of the 4 factors by form version. We see some evidence that the order of the items may affect youth responses. At all points of administration, youth rate themselves lower on negative self-concept items (in other words, they rate themselves more positively) when they are towards the end of the survey (in Form B). However, this difference is not significant at Guatemala endline. At Uganda baseline, we see that youth rate themselves more positively on communication & social skills when these items are at the end of the survey (in Form A). However, we see the opposite effect at Guatemala baseline and endline and Uganda endline—although only the Guatemala baseline finding is significant. Overall, the findings suggest that the order of items may affect how youth rate themselves—however, we recommend further research to test this assumption.

Table 10. Analysis of Youth Skill Levels by Item Order

	Baseline			Endline		
Guatemala	Survey A	Survey B	Diff	Survey A	Survey B	Diff
Positive Self Concept	3.37	3.39	0.02	3.28	3.25	-0.03
Negative Self Concept	2.08	1.99	-0.09*	2.18	2.17	-.001
HOTS	2.99	2.97	-0.02	2.93	2.95	0.02
Communication & Social Skills	2.94	2.99	0.06*	2.91	2.95	0.04
Observations	**399**	**395**		**408**	**376**	
Uganda	Survey A	Survey B	Diff	Survey A	Survey B	Diff
Positive Self Concept (EFA)	4.05	4.23	0.18*	3.47	3.49	0.02
Negative Self Concept (EFA)	2.36	2.30	-0.06*	1.99	1.94	-0.05*
HOTS (EFA)	3.71	3.69	-0.02	3.23	3.24	0.01
Communication & Social Skills (EFA)	3.43	3.25	-0.19*	3.22	3.24	0.02
Observations	**519**	**579**		**572**	**438**	

Notes: Significance is denoted as: * $p < 0.05$

Analysis of Enumerator Characteristics

Our analysis also considers the potential effects of enumerator gender, age, and research experience on youth soft skills and validation outcomes. Our findings suggest that enumerators' gender, age, and research experience all affect youth's responses, although it is unclear why.

Table 11 presents the key descriptive statistics for the 31 enumerators from Uganda and 17 enumerators from Guatemala (endline only). Table 12 presents data on the enumerator characteristics after merging the enumerator datasets with the youth datasets.

Table 11. Descriptive Statistics on Enumerators, Uganda and Guatemala, Endline

	Uganda	Guatemala
Gender		
Female	0.65	0.47
Years of research experience		
0-1 years	0.00	0.24
2-5 years	0.48	0.53
6 or more	0.52	0.24
How often do you do survey research?		
1-2 times a year	0.23	0.12
3-4 times a year	0.58	0.18
5 or more times	0.19	0.71

The respondents who were surveyed by females in Uganda reported a significantly lower positive self-concept score, HOTS score, and communication and social skills score than those surveyed by male surveyors. Corollary to this is a significantly lower negative self-concept score reported by respondents surveyed by females. A significantly higher proportion of respondents reported having some disability to female enumerators. On the other hand, a significantly lower proportion of respondents reported of having sex to a female enumerator. The enumerator gender effect, seen in Table 13, is evident in the time required to conduct the survey, with female enumerators requiring more time to complete the survey. The gender effects are not significant in Guatemala study, except that a significantly lower proportion of respondents reported having had sex to female numerators compared to male enumerators.

Table 12. Enumerator Characteristics after merging the respondent data

	Uganda	Guatemala
Gender		
Male	0.69	0.53
Female	0.31	0.47
Age		
25 years and below	0.11	0.44
Above 25 years	0.89	0.56
35 years and below	0.81	0.93
Above 35 years	0.19	0.07
Research Experience		
0-1 years	0.00	0.29
2 and above	1.00	0.71
0-5 years	0.49	0.88
6 and above	0.51	0.12
Survey Experience		

1-2 times a year	0.24	0.15
3 and more times a year	0.76	0.85
1-4 times a year	0.85	0.35
5 and more times a year	0.15	0.65
Observations	**1010**	**783**

Table 13. Enumerator gender effects on youth responses, item non-response, and survey time

	Uganda		Guatemala	
Gender	Female	Male	Female	Male
Positive Self Concept	3.47	3.51**	3.26	3.28
Negative Self Concept	2.05	1.97***	2.18	2.16
HOTS	3.22	3.27**	2.94	2.96
Communication & Social Skills	3.21	3.25	2.94	2.93
Any Violence	0.68	0.69	0.47	0.50
Any Disability	0.45	0.38*	0.53	0.54
Ever had sex	0.21	0.29***	0.12	0.16
Employment score	0.34	0.33	0.34	0.36*
Missing values				
Any Violence	0.01	0.00	0.04	0.05
Ever had sex	0.00	0.01	0.02	0.05**
Time to complete survey (minutes)	**61.23**	**49.35*** **	**25.96**	**26.99**
Observations	**693**	**317**	**414**	**369**

Asterisks denote statistical significance as follows. *** p<0.01, ** p<0.05, * p<0.1

Table 14 indicates enumerator age effects on participant responses. In Uganda, respondents reported a significantly lower positive self-concept score and a significantly higher negative self-concept score to younger enumerators than to older enumerators. Also, a significantly higher proportion of respondents reported to having had sex to younger enumerators. On an average, younger enumerators took more time to complete the survey than the older enumerators. The interviewer age effects are not significant in Guatemala, except that respondents report a higher negative self-concept score to younger enumerators. This result is consistent with the interviewer effect in Uganda study.

Table 14. Enumerator age effect on youth responses, item non-response and survey time

	Uganda		Guatemala	
Age	35 years and below	Above 35 years	35 years and below	Above 35 years
Positive Self Concept	3.47	3.55***	3.27	3.31
Negative Self Concept	2.05	1.93***	2.18	2.05***
HOTS	3.23	3.29**	2.94	2.96
Communication & Social Skills	3.22	3.26	2.93	2.91
Any Violence	0.68	0.68	0.49	0.43
Any Disability	0.43	0.43	0.53	0.55
Ever had sex	0.25	0.16**	0.14	0.15
Employment score	0.34	0.30**	0.35	0.33

Missing Value				
Any Violence	0.00	0.01	0.05	0.00
Ever had sex	0.00	0.01	0.03	0.02
Time	**56.67**	**60.98**	**26.57**	**24.81**
Observations	**821**	**189**	**730**	**53**

Table 14 indicates the effect of interviewer's research experience on the participant responses. In Uganda, the respondents reported a significantly higher negative self-concept score, employment score and incidents of violence to enumerators with less than 6 years of research experience than to those with 6 or more years of experience. Also, enumerators with 2-5 years of research experience took less time to complete the survey than the more experienced enumerators. In contrast, the Guatemala study indicates that the respondents report a lower negative self-concept score to enumerators with less research experience and the enumerators with less research experience take significantly more time to complete the survey.

Table 15. Enumerator research experience effect on youth responses, item non-responses, and survey time

	Uganda		**Guatemala**	
Research Experience	2-5 years	6+	0-5 years	6+
Positive Self Concept	3.48	3.48	3.27	3.24
Negative Self Concept	2.05	2.00**	2.16	2.26***
HOTS	3.24	3.23	2.94	2.96
Communication & Social Skills	3.24	3.22	2.93	2.94
Any Violence	0.71	0.65**	0.50	0.41
Any Disability	0.45	0.41	0.54	0.52
Ever had sex	0.22	0.25	0.14	0.09
Employment score	0.37	0.31***	0.35	0.33
Missing Value				
Any Violence	0.00	0.01	0.05	0.02
Ever had sex	0.00	0.01	0.03	0.03
Time	**51.36**	**63.27*** **	**27.18**	**21.05*** **
Observations	**492**	**518**	**690**	**93**

In sum, these findings demonstrate that enumerator gender, age, and research experience may affect how youth respond to survey questions, and that these effects can differ by country context.

Conclusion

Results from two validation sites indicate that **the instrument is internally consistent and valid**, as demonstrated through both the Cronbach's alpha statistics, test-retest statistics, and correlation coefficients between soft skills and external characteristics and outcomes. We also find that the tool is invariant by country, time point, gender, and SES. Despite identifying some items as having DIF, the magnitude of the differences in those items do not in any case lead to noticeable differences at the construct level. Finally, we find that the tool measures change over time; however, the change is not in the expected direction, and we cannot attribute this change to the partner programs.

The originally hypothesized structure changed somewhat during the process of the tool development—negative self-concept emerged as a construct distinct from positive self-concept, and social skills and

communication skills clustered together. **Our empirically determined scale structure demonstrates a good fit with the data.**

Validation of the tool revealed several interesting findings. **First, SES and disability status are both highly predictive of soft skills.** Higher SES was associated with higher positive self-concept and higher order thinking skills in Uganda and higher positive self-concept and lower negative self-concept (significant only at endline) in Guatemala. Youth who report having any disability report lower levels of all skills (and higher negative self-concept) in both countries. **Employment status, violent behavior, and to some extent, RH behaviors, are also predictive of SES.** Youth's employment outcomes are associated with small but significant changes in the expected direction for all skills (except for negative self-concept in Uganda). Youth who reported having engaged in any kind of violence also report lower levels of skills across the board (and higher levels of negative self-concept). RH behaviors are significantly associated with lower positive self-concept and higher negative self-concept in Uganda and *higher* social and communication skills in Guatemala.

In addition to these psychometric tests, we performed several other analyses that revealed useful information for future development of soft skills measures. Our integration of anchoring vignettes into the tool shows **the utility of anchoring vignettes as a technique for assessing respondents' engagement with and comprehension of an instrument**. However, because the AV-adjusted scores did not improve the scales, we recommend the use of the non-adjusted scores for the main analysis, except in the case of cross-country comparison.

Our inclusion of a program staff tool showed that **a third-party assessment may not be an analytically useful way of capturing youth's skill levels.** Our analysis of correlations between the youth and program staff factors shows generally low correlations for both countries and at both times, although we see some changes in the expected direction when we compare the endline and baseline correlations. Some of the program staff tool subscales seem to hold together reliably, but overall, we find that the results of the youth and program staff data together are neither easily interpretable nor analytically useful.

In addition, our **analysis of enumerator characteristics shows that they do matter**. In Uganda, and in a few cases for Guatemala, we see effects on youth's responses by enumerators' gender and age. Enumerators' years of research experience also has an effect on youth responses for several variables in Uganda and one variable in Guatemala.

In sum, the YouthPower Action Youth Soft Skills measurement tool presents a validated assessment of youth soft skills such as positive self-concept, negative self-concept, higher order thinking skills, and social and communication skills. The assessment can be used to predict youth outcomes in key areas and measure change in the level of soft skills over time, in as short as a few months. Further testing is necessary to determine validity in contexts substantially different than the youth programs in which the assessment was validated, with other youth outcomes, and in a causal inference framework.

Recommendations

For Tool Development

Our experience in developing this instrument reveals several key lessons.

1) **Contextualization is a critical first step**. The cognitive testing that we conducted in Uganda and Guatemala revealed critical information—especially in Uganda where we underwent a lengthier process—that informed our initial revisions and helped us to interpret several confounding findings. For example, data from the cognitive interviews in Uganda showed that

items measuring delayed gratification were not necessarily contextually appropriate. This informed our decision to ultimately remove these items from the instrument.

2) **Less is more.** Our analysis of the Uganda baseline data showed that the most low-performing items were often also the wordiest. Overall, the tool performed better once we radically simplified the item wording and structure. Because youth were not using all 5 response options provided to them in the first version of the tool, that we could reduce the number of response options to 4.
   We also found that additional tools may not provide additional analytical clarity. Specifically, our analysis of data from a program staff, or observer, tool revealed extremely low correlations between youth's and program staff's assessments of youth skills, suggesting the limited utility of the observer report.

For Implementation

3) Implementers seeking to measure program participants' soft skills levels and progress should build in sufficient time and resources in the program cycle to conduct baseline (and endline, where relevant) assessments—and to contextualize the assessment for the program and country context. The time required to administer a survey will depend primarily on the readiness of the tool for administration (has it been tested in this context with this particular demographic of youth before?); the number and geographic spread of youth to be surveyed; and the number of enumerators.
4) Implementers should also take a wide view of monitoring and evaluation processes **to ensure that participants do not experience survey fatigue**. Our finding regarding Form A versus Form B revealed that respondents may respond differently towards the end of surveys, possibly due to fatigue. Thus, ensuring that no one participant takes too many surveys, and that no survey is too long (what is "too long" will differ depending on the individual, but we suggest no longer than 45 minutes and, ideally, closer to 20 minutes) is critical in collecting quality data.
5)

For Research

We recommend that future research studies focus on the following gaps:

6) **Testing of this tool in the context of an aligned soft skills intervention, with a control group** in order to more clearly isolate where change in skill levels may be coming from. Future studies could also focus on differences in skill levels by intervention dosage and/or modality in order to discern how much of an intervention is required to "move the needle" on a certain skill, and whether certain intervention modalities work better than others, especially by cultural context.
7) **Further research on the pathways linking skills and outcomes.** Our analysis found that higher communication and social skills among youth were linked to a higher likelihood that youth had ever had sex in Guatemala. Our earlier literature review (Gates et al., 2016) found several studies linking higher skill levels to violent behavior and risky sexual behaviors. Research exploring the relationship between youth soft skills and specific employment outcomes, such as on-the-job performance, retention, and wages, would also be useful for understanding how youth's skills support them in the workforce.
8) **Testing and validation of this tool in additional contexts** in order to better understand cross-cultural differences in how skills change over time, how skills relate to key outcomes, and how skills relate to other key demographics, such as SES, gender, and disability—and whether other environmental factors need to be taken into consideration when planning soft skills interventions**.** Testing of the tool in other countries may also reveal different response styles among youth, in which case the AVs might be applied in order to compare data across contexts.

9) **Further research should be conducted on the effect of enumerator gender, age, and research experience** in different contexts. Our research revealed that these characteristics can have an effect on youth responses. Why and how they have an effect is less clear and warrants further research.

References

Almagor, M., Tellegen, A., & Waller, N. G. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. Journal of Personality and Social Psychology, 69(2), 300-307.http://dx.doi.org/10.1037/0022-3514.69.2.300

An, S., Ji, L.-J., Marks, M., & Zhang, Z. (2017). Two Sides of Emotion: Exploring Positivity and Negativity in Six Basic Emotions across Cultures. Frontiers in Psychology, 8, 610. DOI=10.3389/fpsyg.2017.00610

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.

Bollen, K. A. (1989). Structural equations with latent variables. New York, NY: Wiley.

Bollen, K. A., & Long, J. S. (Eds.). (1983). Testing structural equation models. Newbury Park, CA: Sage.

Card, N. A. (2016). Methodological issues in measuring the development of character. *Journal of Moral Education*.

Center for the Economics of Human Development. (2015). Conference on Measuring and Assessing Skills Report. Chicago: University of Chicago Chen, F. F. (2007) Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance, Structural Equation Modeling: A Multidisciplinary Journal, 14:3, 464-504, DOI: 10.1080/10705510701301834.

Connor-Smith, J. K., Compas, B. E., Wadsworth, M. E., Thomsen, A. H., & Saltzman, H. (2000). Responses to stress in adolescence: measurement of coping and involuntary stress responses. *Journal of consulting and clinical psychology*, *68*(6), 976.

Cronbach, L. J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper & Row.

Davier, M., Shin, H., Khorramdel, L., and Stankov, L. (2017). The effects of vignette scoring on reliability and validity of self-reports. Applied Psychological Measurement 42(4):291–306.

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.

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. Journal of Cross-Cultural Psychology, 48, 319-334.

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

Hoyle, R. H. (1995). Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA: Sage Publications.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1e55.

King, G., & Lau, O. (2008). Anchors: software for anchoring vignette data. Journal of statistical software.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. Political Analysis, 15, 46–66.

Kyllonen, P. C., & Bertling, J. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 277-286). 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.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 525–543.

Noguchi, K., Gohm, C. L., Dalsky, D. J., Sakamoto, S. (2007). Cultural differences related to positive and negative valence. Asian Journal of Social Psychology 10(2):68 – 76. DOI: 10.1111/j.1467-839X.2007.00213.x

Primi, R., Zanon, C., Santos, D., De Fruit, F. & John, O. P. (2016). Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? European Journal of Psychological Assessment, 32, 39–51.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1 - 36. doi:http://dx.doi.org/10.18637/jss.v048.i02

Simms, L. J. (2007). The Big Seven model of personality and its relevance to personality pathology. Journal of Personality, 75(1), 1-18. https://pdfs.semanticscholar.org/19d5/ba23b50c86d0c46895664bb2cd49b09c993e.pdf

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.

Stone, L. L., Janssens, J. M., Vermulst, A. A., Van Der Maten, M., Engels, R. C., & Otten, R. (2015). The Strengths and Difficulties Questionnaire: psychometric properties of the parent and teacher version in children aged 4–7. *BMC psychology*, *3*(1), 4.

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs (pp. 35- 37). Windsor, England: NFER-NELSON.

Tellegen, A., & Waller, N. G. (1987, August). Re-examining basic dimensions of natural language trait descriptors. In *95th Annual Convention of the American Psychological Association, New York*.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 2, 4–69.

Watson, David, Clark, Lee A., Tellegen, Auke (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology. 54 (6), 1063–1070. doi:10.1037/0022-3514.54.6.1063

Appendix Materials

Appendix 1. Five-Factor versus Four-Factor Structure of Soft Skills Tool

Table A. 1 Five-Factor versus Four-Factor Structure of Soft Skills Tool

Item	Factor according to 5-factor theoretical structure	Decision	Revised Item	Factor according to 4-factor EFA-determined structure	Decision
How often do you believe that: it is hard for you to solve your problems?	Positive self-concept	dropped due to low loading	no equivalent	no equivalent	n/a
How often do you believe that: there are many things that you do poorly?	Positive self-concept	revise	There are many things that I can't do very well	Negative self-concept	keep
How often do you believe that: you are good at learning something new?	Positive self-concept	revise	I'm good at learning new things	Positive self-concept	keep
How often do you believe that: you can do most things if you make the effort?	Positive self-concept	revise	I can do most things if I make an effort	Positive self-concept	keep
How often do you believe that: you can do something that will help you succeed in life?	Positive self-concept	revise	I can do things that will help me succeed in life	Positive self-concept	keep
How often do you feel that you are not good at all?	Positive self-concept	revise	I think I am no good at all	Negative self-concept	keep
How often have you felt that: the people you live with at home value you?	Positive self-concept	revise	I feel valued by the people I live with at home	Positive self-concept	keep
How often do you feel that you are a valued member of your community?	Positive self-concept	revise	I'm a valued member of my community	Positive self-concept	keep

How often have you felt that: you have a number of good qualities?	Positive self-concept	revise	I have a number of good qualities	Positive self-concept	keep
How important is this to you: Liking yourself just the way you are?	Positive self-concept	revise	I like myself just the way I am	Positive self-concept	keep
How often do you feel good about your skills?	Positive self-concept	revise	I feel good about my skills	Positive self-concept	keep
How often do you feel not sure that you can be successful?	Positive self-concept	revise	I'm not sure I can be successful	Negative self-concept	keep
How often do you feel that you don't trust your skills?	Positive self-concept	revise	I'm not confident about my skills	Negative self-concept	keep
How often do you feel confident in yourself?	Positive self-concept	revise	I feel confident in myself	Positive self-concept	keep
How often do you find it hard to know how you are feeling?	Positive self-concept	revise	It is hard to know what I'm feeling	Negative self-concept	keep
How often do you know what you are good at?	Positive self-concept	revise	I know what I'm good at	Positive self-concept	keep
How often do you know how you are feeling inside at any particular moment?	Positive self-concept	revise	I know how I'm feeling inside at any particular moment	Positive self-concept	keep
How often do you know how you make other people feel?	Positive self-concept	revise	dropped due to low loading	no equivalent	n/a
How often do you see that your future will be happy?	Positive self-concept	revise	My future will be happy	Positive self-concept	keep
How often do you believe that you will reach your future goals?	Positive self-concept	revise	I can achieve most of my future goals	Positive self-concept	keep
How often do you know that you are going to be fine?	Positive self-concept	revise	I know I'm going to be fine	Positive self-concept	keep

How often do you believe you can make things happen that will improve your life?	Positive self-concept	revise	I can make things happen that will improve my life	Positive self-concept	keep
How often do you save your money for something you want to buy later?	Self-control	dropped due to low loading	no equivalent	no equivalent	n/a
How often do you find it challenging to wait for something?	Self-control	dropped due to low loading	no equivalent	no equivalent	n/a
How often would you prefer to get one pen now rather than many pens later?	Self-control	dropped due to low loading	no equivalent	no equivalent	n/a
How often do you do things without thinking about what you're doing?	Self-control	revise	I do things before I think through them	Negative self-concept	keep
In the past month, how often have you interrupted your friend when they were telling a story?	Self-control	revise	dropped due to low loading	no equivalent	n/a
How often do you think through things before you do them?	Self-control	revise	I think carefully before doing anything	HOTS	keep
In the past month, how often have you finished the work that you set out to do despite challenges?	Self-control	revise	dropped due to low loading	no equivalent	n/a
In the past month, how often have you been unable to pay attention?	Self-control	revise	I have a hard time concentrating on one thing.	Negative self-concept	keep
In the past month, how often have you kept doing something that you should do even if you didn't like it, such as homework?	Self-control	dropped due to low loading	no equivalent	no equivalent	n/a

In the past month, how often have you found it difficult to start your work?	Self-control	revise	I have difficulty starting tasks	Negative self-concept	keep
In the past month, how often have you done things to control your anger or temper, for example when you have quarreled with your friend?	Self-control	revise	When things go wrong for me, I'm good controlling my temper	HOTS	keep
In the past month, how often have you been annoyed by little things, like if someone steps on your shoe?	Self-control	revise	I'm easily annoyed by little things (like if someone steps on my shoe)	Negative self-concept	keep
In the past month, how often have you remained calm when a friend tells you that you did something poorly?	Self-control	revise	If a friend tells me I did something wrong, I can stay calm	n/a	Dropped due to low loading
In the past month, how often were you able to stop yourself when you were going to do something you would regret?	Self-control	revise	If I'm doing something that I know I would regret, I'm able to stop before it is too late	Positive self-concept	keep
In the past month, how often have you refused to follow instructions?	Self-control	revise	I'm good at following instructions	Social and communication skills	keep
In the past month, how often have you got your work done immediately instead of waiting until the last minute?	Self-control	dropped due to low loading	no equivalent	no equivalent	n/a
How often do you do crazy things, such as drinking alcohol, even if they are a little dangerous?	Self-control	dropped due to low loading	no equivalent	no equivalent	n/a
How often do you do what feels good to you without thinking about its results?	Self-control	revise	I do whatever feels good to me, without thinking about the results	Negative self-concept	keep

How often do you do something risky because of peer pressure?	Self-control	revise	If my friends are doing something risky, I will do it with them	Negative self-concept	keep
When answering these next four questions, think about the last few problems you have had in the past month, like when an object breaks			When answering these next four questions, think about the last few problems you have had and tell us how much you agree with each statement.		
In the past month, how often did you take action to solve the problems?	Higher order thinking skills	revise	I took action to solve the problems	HOTS	keep
In the past month, how often did you ask other people for help with the problems?	Higher order thinking skills	revise	I asked other people for help to solve the problems	HOTS	keep
In the past month, how often did you try to think of different ways to solve the problems?	Higher order thinking skills	revise	I tried to think of different ways to solve the problems	HOTS	keep
In the past month, how often did you make a plan to solve the problems?	Higher order thinking skills	revise	I made a plan to solve the problems	HOTS	keep
When answering these next three questions, think about the last few times someone told you an interesting story			When answering these next three questions, think about the last few times someone told you an interesting story and tell us how much you agree with each statement.		
How often did you separate the true and false parts of the story?	Higher order thinking skills	dropped due to low loading	no equivalent	no equivalent	n/a
How often did you question why someone in the story did what they did?	Higher order thinking skills	revise	I questioned why someone in the story did what they did	HOTS	keep

How often did you connect pieces of evidence together?	Higher order thinking skills	revise	I connected pieces of evidence together	HOTS	keep
When answering these next three questions, think about the last few times you made a decision.			When answering these next three questions, think about the last few times you made a decision and tell us how much you agree with each statement.		keep
Before making the decisions, how often did you collect a lot of information?	Higher order thinking skills	revise	I collected a lot of information before making the decision	HOTS	keep
Before making the decisions, how often did you think about how others would be affected?	Higher order thinking skills	revise	I thought about how other people would be affected before making the decision	HOTS	keep
Before making the decisions, how often did you consider different options?	Higher order thinking skills	revise	I considered different options before making the decision	HOTS	keep
How often do you avoid making your friends look bad?	Social skills	revise	I'm able to stand up for myself without putting others down	Social and communication skills	keep
How often do you find a way to work things out if two of your friends quarrel?	Social skills	revise	I find a way of working things out if two of my friends quarrel	n/a	Dropped due to low loading
How often do you do your part when you work in a group?	Social skills	dropped due to low loading	no equivalent	no equivalent	n/a
How often do you relate well with people of different backgrounds?	Social skills	revise	I get along well with people from different backgrounds	HOTS	keep
How often do you find it easy to make friends?	Social skills	revise	I find it easy to make friends	HOTS	keep

How often do you control your anger when you have a misunderstanding with a friend?	Social skills	revise	I can control my anger when I have a misunderstanding with a friend	Social and communication skills	keep
How often do you respect views that differ from your own?	Social skills	revise	dropped due to low loading	no equivalent	n/a
How often do you write a story or letter well?	Communication	revise	I write well.	Social and communication skills	keep
How often can you discuss a problem with a friend without making things worse?	Communication	revise	I am good at resolving disagreements.	Social and communication skills	keep
How often are you uncomfortable to ask questions in class?	Communication	revise	It is easy for me to ask questions in a public setting.	Social and communication skills	keep
How often are you rude to others?	Communication	revise	I am rude to others.	Negative self-concept	keep
How often do you tell others how you feel?	Communication	revise	It is easy for me to share my feelings with others.	Social and communication skills	keep

Appendix 2. Scale Sample Composition and Descriptive Statistics

Sample Composition

Sample statistics are presented in **Error! Reference source not found.** below. The samples were well balanced at baseline and endline and tended to include more female participants in both countries. The Uganda participants were generally somewhat older on average than the Guatemala participants. Due to the size of the programs, the Uganda sample size was larger than that in Guatemala.

Table A. 2 Sample Statistics, Uganda and Guatemala, Baseline and Endline

	Uganda					
	All Sample			Panel Sample		
	Baseline n=1,089	Endline n=1,010	Diff (E-B)	Baseline n=556	Endline n=556	Diff (E-B)
Demographics						
Female	0.53	0.54	0.01	0.57	0.55	-0.02
Age	17.05	17.34	0.29*	17.13	17.42	0.29*
SES Index	-0.02	0.02	0.04	0.08	0.01	-0.07
Interview conducted in English	.86	.96	.09*	.89	.96	.08*
Educate! scholar	.76	.76	.01	.75	.87	.12*

	Guatemala					
	All Sample			Panel Sample		
	Baseline n=794	Endline n=784	Diff (E-B)	Baseline n=450	Endline n=450	Diff (E-B)
Demographics						
Female	0.53	0.57	0.05	0.56	0.58	0.01
Age	16.42	16.83	0.40*	16.44	16.76	0.32*
SES Index	0.05	-0.05	-0.10*	-0.02	-0.09	-0.07
Spanish spoken at home	.69	.67	-.02	0.69	.72	.03

Note: Significance is denoted as: * $p < 0.05$. Panel Sample contains only the youth who were interviewed both at baseline and endline. All Sample contains all the youth interviewed in each time period.

Descriptive Statistics by Skill

Error! Reference source not found. shows descriptive statistics, or mean scores, for each scale, for Uganda and Guatemala, at baseline and endline. Note that, following the analysis of Uganda baseline administration data, the team made significant revisions to the tool and ultimately adopted a different (four-) factor structure from the originally proposed (five-factor) theoretical structure. This new factor structure was then used to analyze data from both countries at both points in time and is reflected below.

Scores for Uganda baseline are not comparable to Uganda endline or the Guatemala data, since they were calculated out of a five-point scale. Uganda endline respondents showed a higher average on positive-self-concept, HOTS, and communication and social skills, and a lower average on negative self-concept skills when compared to Guatemala. Guatemala endline respondents showed a lower average

on positive self-concept, HOTS, and social communication skills, and a higher average on negative self-concept when compared to baseline respondents.

Table A. 3 Descriptive statistics Uganda and Guatemala, Baseline and Endline

	Uganda		Guatemala	
	Baseline*	Endline	Baseline*	Endline
Positive Self Concept	4.15	3.48	3.38	3.27
Negative Self Concept	2.33	1.97	2.04	2.17
HOTS	3.70	3.24	2.98	2.94
Communication & Social Skills	3.33	3.23	2.97	2.93
Observations	**1098**	**1010**	**794**	**784**

Response Option Use and Patterns

Error! Reference source not found. shows the average frequency of response option by scale and time period, for Uganda and Guatemala. This helps us understand respondents' response patterns—for example, whether they only used choices at the "extreme" ends of the scale or whether their responses were distributed more evenly. The response patterns differ depending on the skill and country and suggest that respondents are more likely to respond using the "extreme" options (never or almost never; always or almost always; strongly disagree; strongly agree) for positive and negative self-concept. For communication and social skills, they are more likely to respond somewhere in the middle, or on the high end of the scale.[11]

Table A. 4 Response Option Frequency, Uganda and Guatemala, Baseline and Endline

Baseline					Endline				
Scales	Pos. Self-Con.	Neg. Self-Conc.	HOTS	Soc. & Com. Skills	Scales	Pos. Self-Con.	Neg. Self-Con.	HOTS	Soc. & Com. Skills
Uganda									
Never or almost never	0.02	0.26	0.04	0.1	Strongly Disagree	0.01	0.28	0.02	0.02
Rarely	0.06	0.31	0.11	0.16	Disagree	0.02	0.46	0.08	0.1
Sometimes	0.21	0.32	0.32	0.32	Agree	0.45	0.2	0.54	0.52
Often	0.17	0.05	0.16	0.13	Strongly Agree	0.52	0.06	0.36	0.36
Almost always or always	0.54	0.06	0.36	0.29					
Guatemala									
Strongly Disagree	0.01	0.2	0.03	0.02	Strongly Disagree	0.01	0.13	0.02	0.02
Disagree	0.04	0.59	0.17	0.21	Disagree	0.06	0.59	0.19	0.21
Agree	0.51	0.18	0.6	0.54	Agree	0.6	0.24	0.64	0.6

[11] **Note** that at baseline in Uganda, the respondents were presented with ***five*** response options; the item stem and response options were subsequently revised going forward so that respondents only had ***four*** response options.

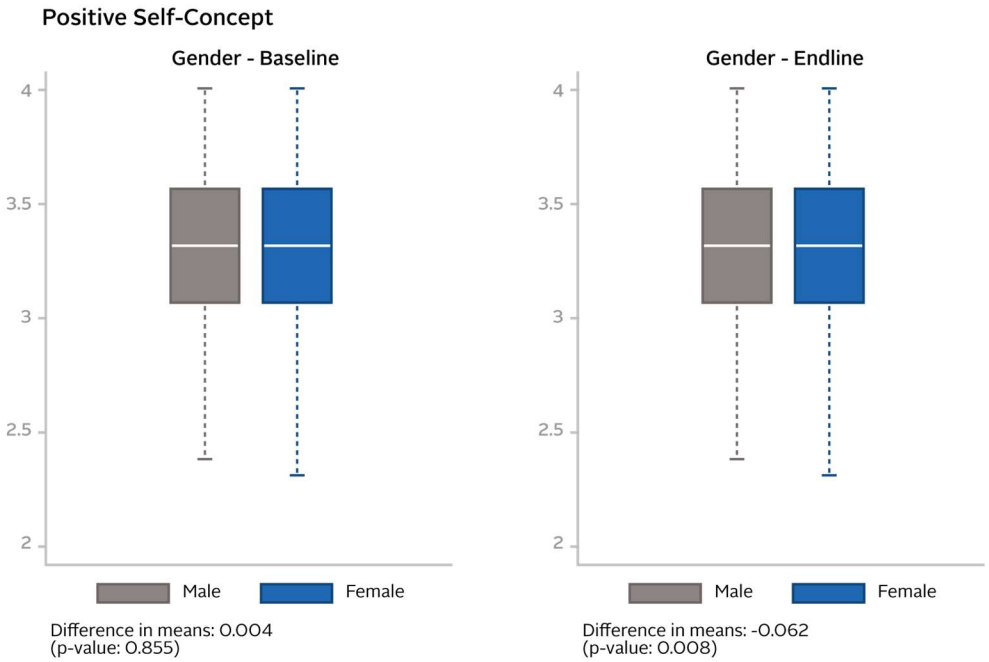
Strongly Agree	0.44	0.03	0.21	0.22	Strongly Agree	0.34	0.03	0.16	0.17

Descriptive Statistics by Skill and Group

Figure A. 1 – A.24 show the box plots for each scale by gender, socio-economic status, rural/urban location, and language used at home. These box plots help us to understand whether responses differ for a particular skill by a certain trait (for example, gender), and whether this pattern is different at baseline versus endline.

Figure A. 1 – A.8 show positive self-concept scores by gender at baseline and endline in Guatemala and Uganda.[12] In Guatemala, males and females show similar distribution at baseline and endline, while females show higher scores and larger variation for negative self-concept at baseline and endline; this difference in means is only significant at endline. Females show lower values and large variation for the HOTS scale although the difference in means is not significant. Social and communication skills scoring is larger for males and the difference in means is only significant at endline. In Uganda, males rate themselves higher on positive self-concept, while females rate themselves higher on negative self-concept (however, neither differences are significant). They rate themselves similarly on HOTS. Females rate themselves lower on social and communication skills, with this difference being significant at endline.

Figure A. 1. Box-plot by gender, Guatemala, Baseline and Endline: Positive Self-Concept



[12] Note that the baseline instrument that was administered in Uganda differs from the endline instrument due to the substantial revisions that were made after analysis of the Uganda baseline data.

Figure A. 2 Box-plot by gender, Uganda, Baseline and Endline: Positive Self-Concept

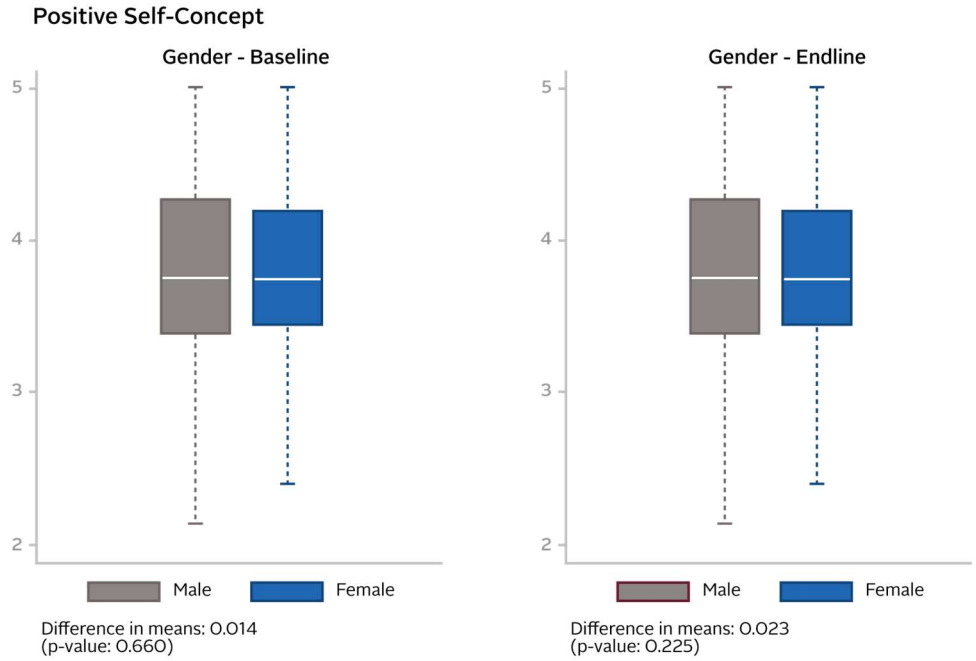


Figure A. 3 Box-plot by gender, Guatemala, Baseline and Endline: Negative Self-Concept

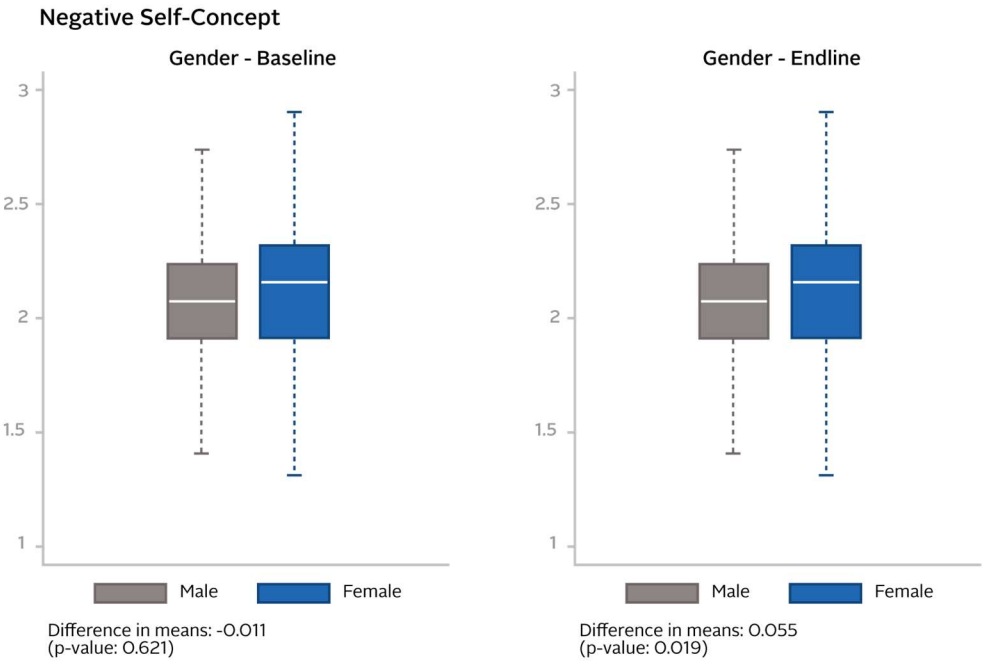


Figure A. 4 Box-plot by gender, Uganda, Baseline and Endline: Negative Self-Concept

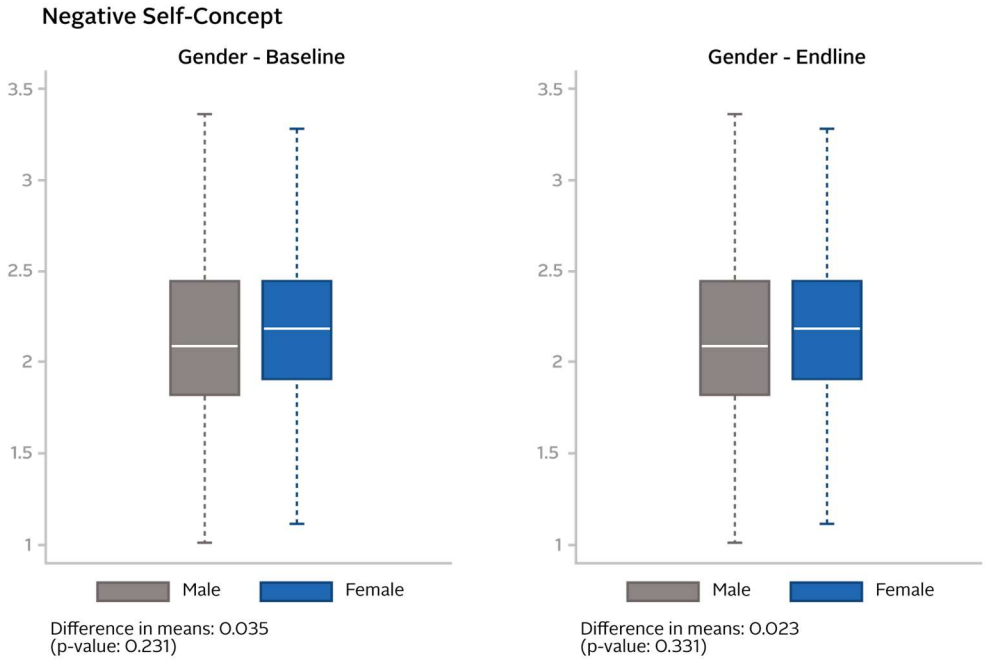


Figure A. 5 Box-plot by gender, Guatemala, Baseline and Endline: HOTS

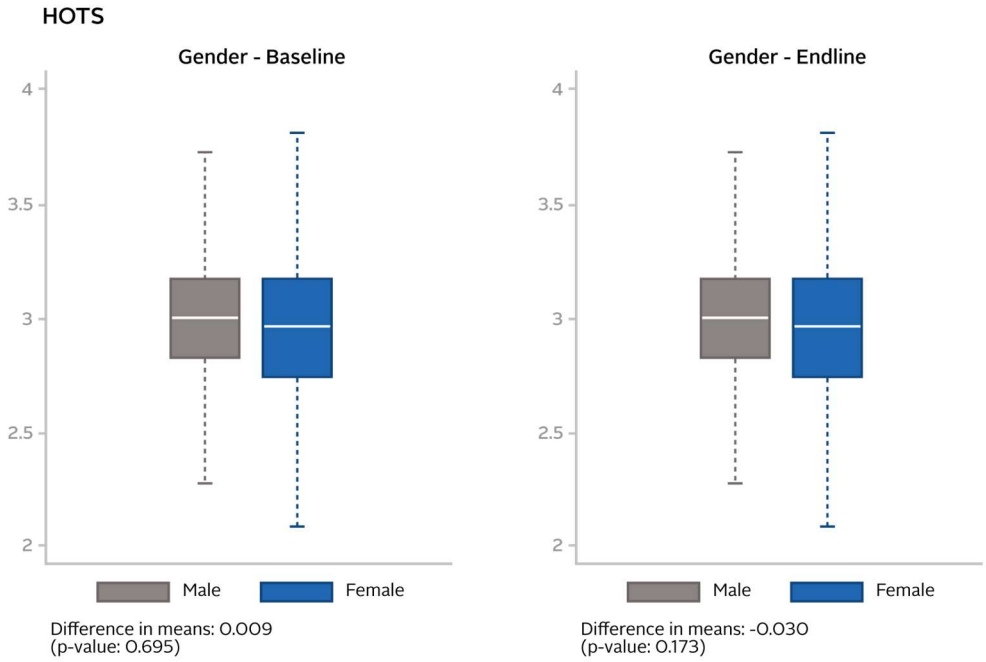


Figure A. 6 Box-plot by gender, Uganda, Baseline and Endline: HOTS

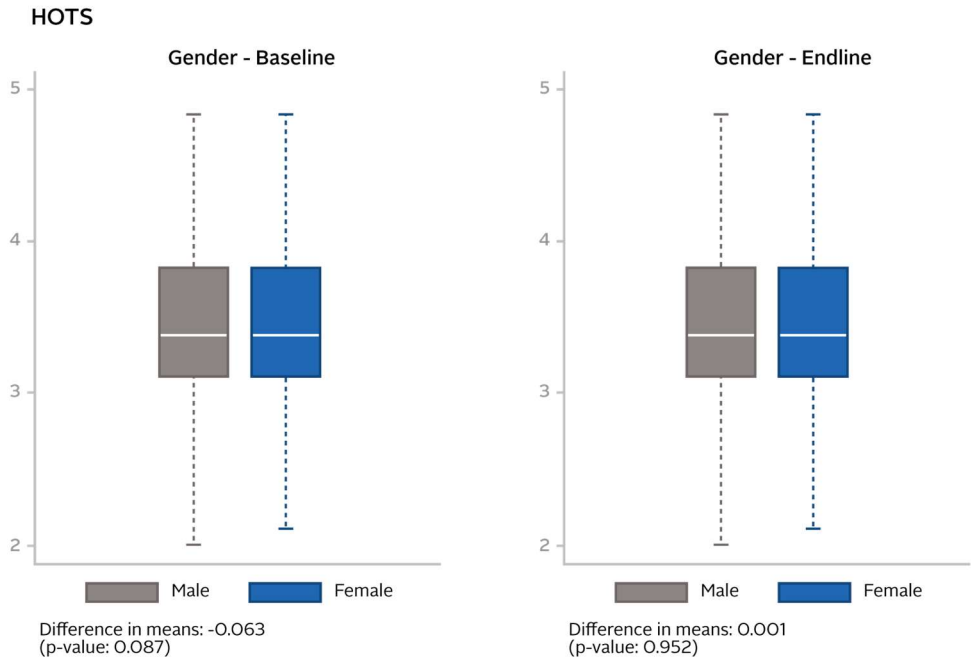


Figure A. 7 Box-plot by gender, Guatemala, Baseline and Endline: Social and Communication Skills

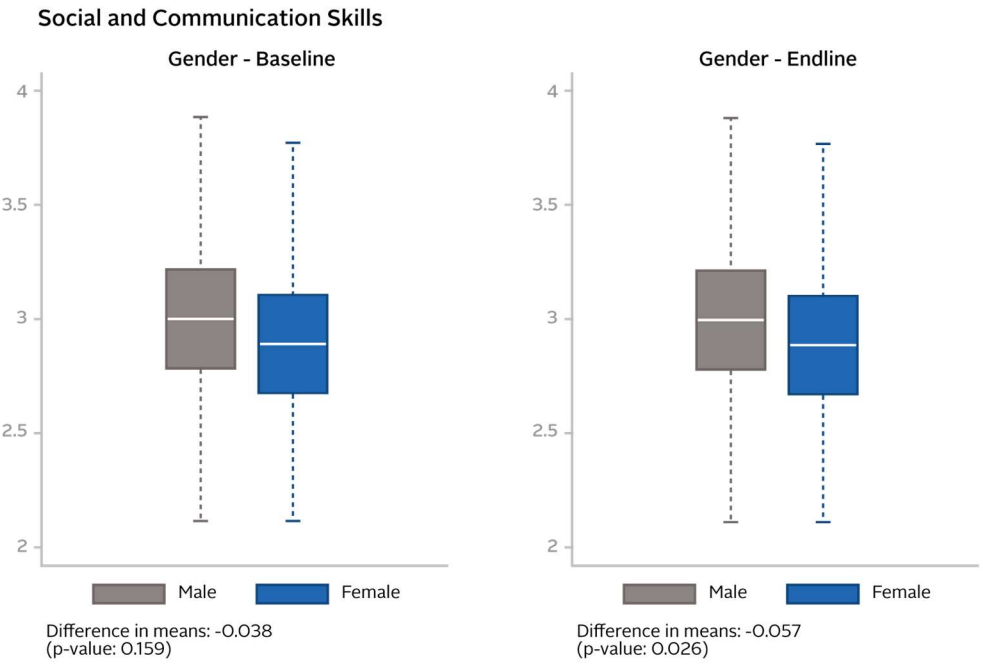
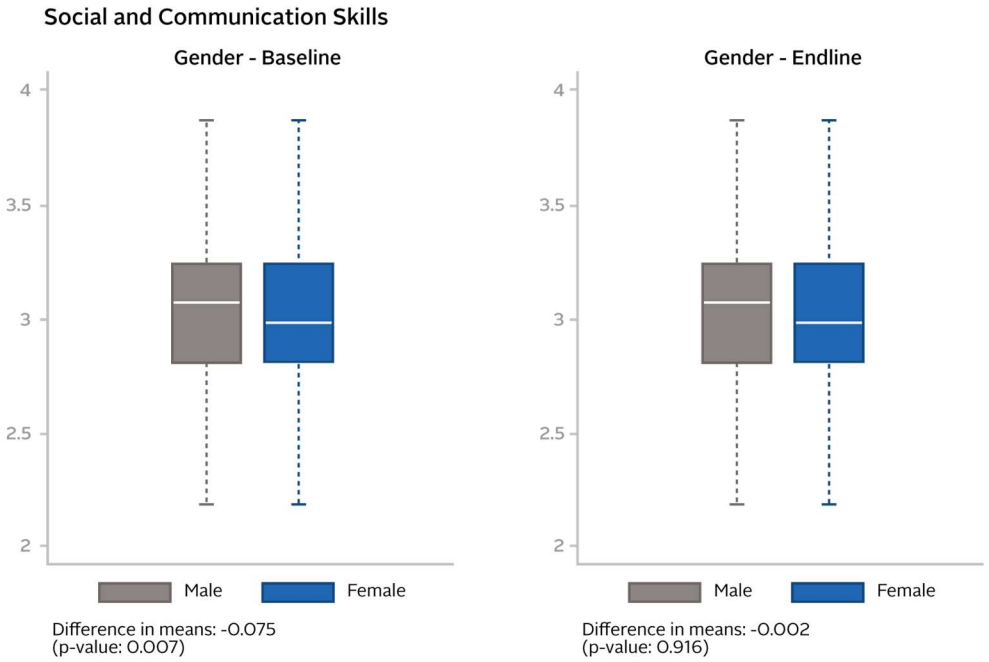


Figure A. 8 Box-plot by gender, Uganda, Baseline and Endline: Social and Communication Skills



Descriptive statistics by low and high SES are shown in A.9 – A.16. In Guatemala, youth with a high SES show higher scoring and larger variation for the positive self-concept scale (difference in means significant only at endline); lower scoring and larger variation for negative self-concept (difference in means significant at both periods); higher median and similar overall distribution for HOTS (difference in means not significant); and larger variation for social and communication skills with a higher concentration of youth scoring above the median (difference in means not significant). In Uganda, youth rate themselves lower on positive self-concept, HOTS, and social and communication skills; these differences are significant for positive self-concept at baseline and endline and for HOTS at baseline and endline. Similarly, youth rate themselves higher on negative self-concept, but the difference in means is not significant.

Figure A. 9 Box-plot by SES, Guatemala, Baseline and Endline: Positive Self-Concept

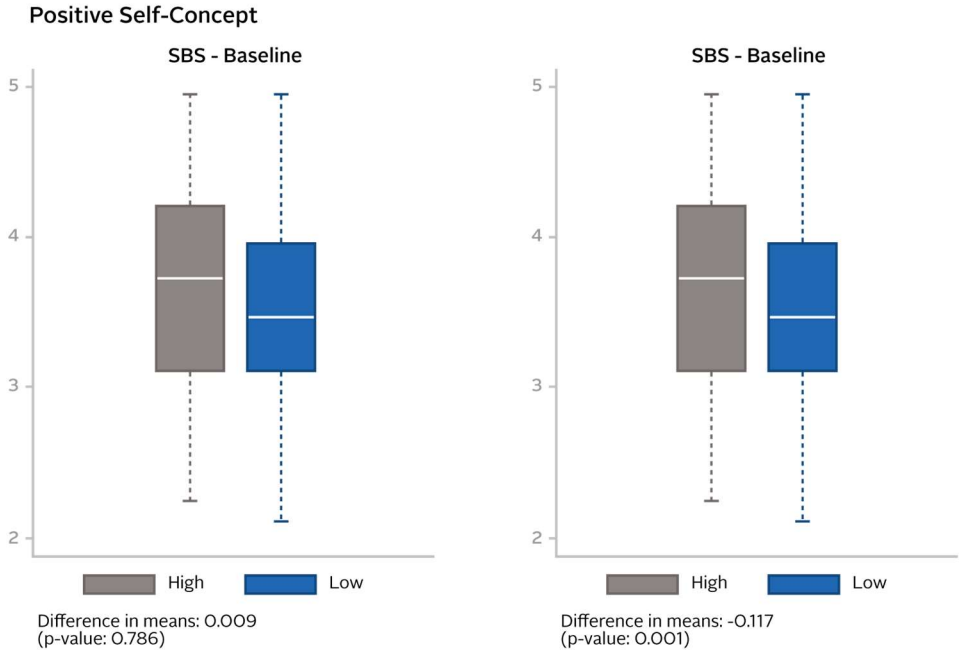


Figure A. 10 Box-plot by SES, Uganda, Baseline and Endline: Positive Self-Concept

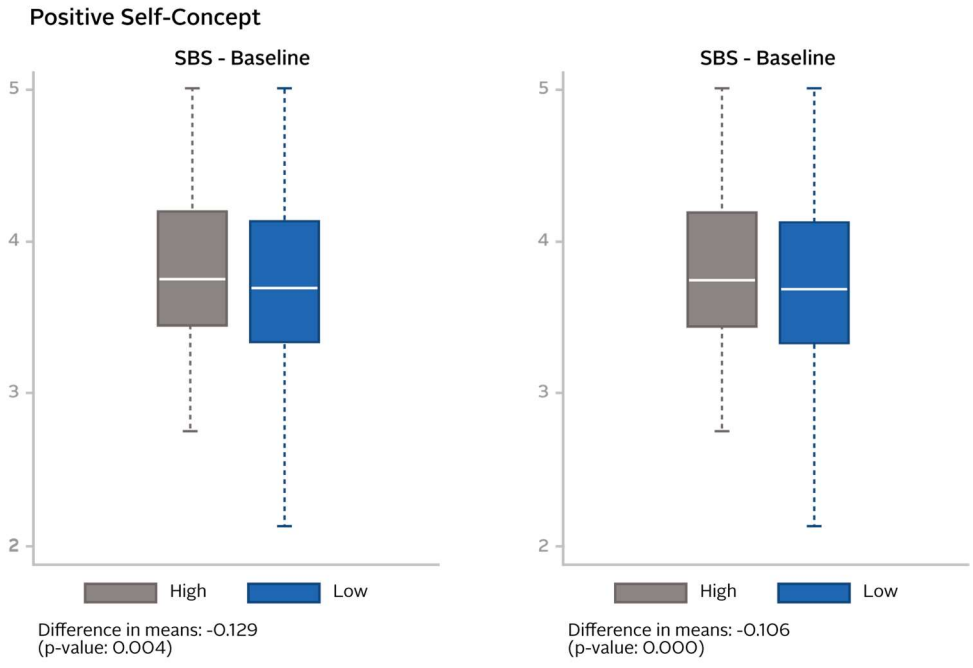


Figure A. 11 Box-plot by SES, Guatemala, Baseline and Endline: Negative Self-Concept

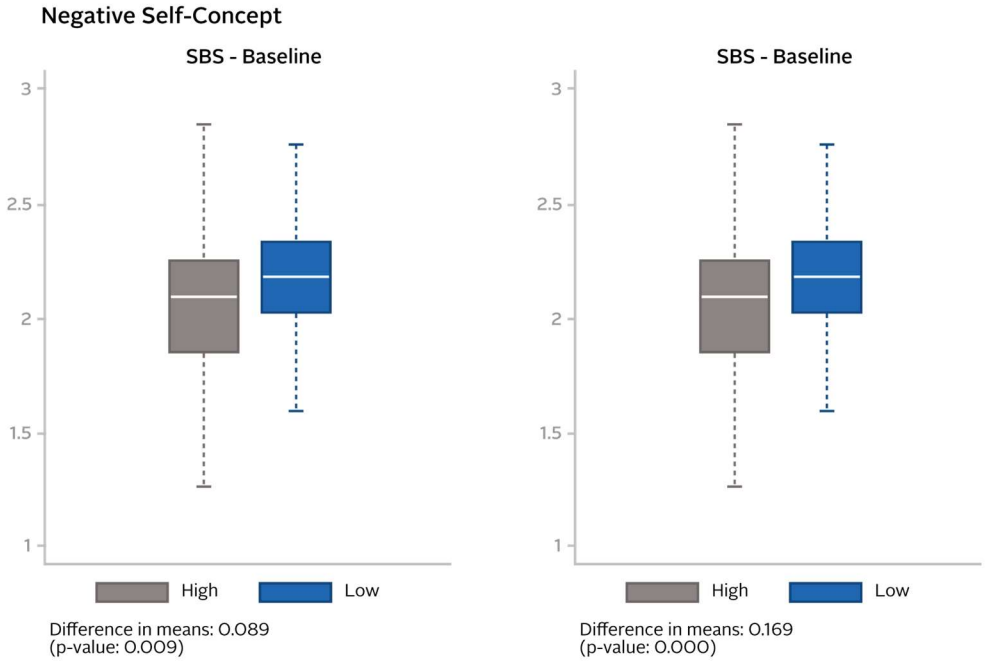


Figure A. 12 Box-plot by SES, Uganda, Baseline and Endline: Negative Self-Concept

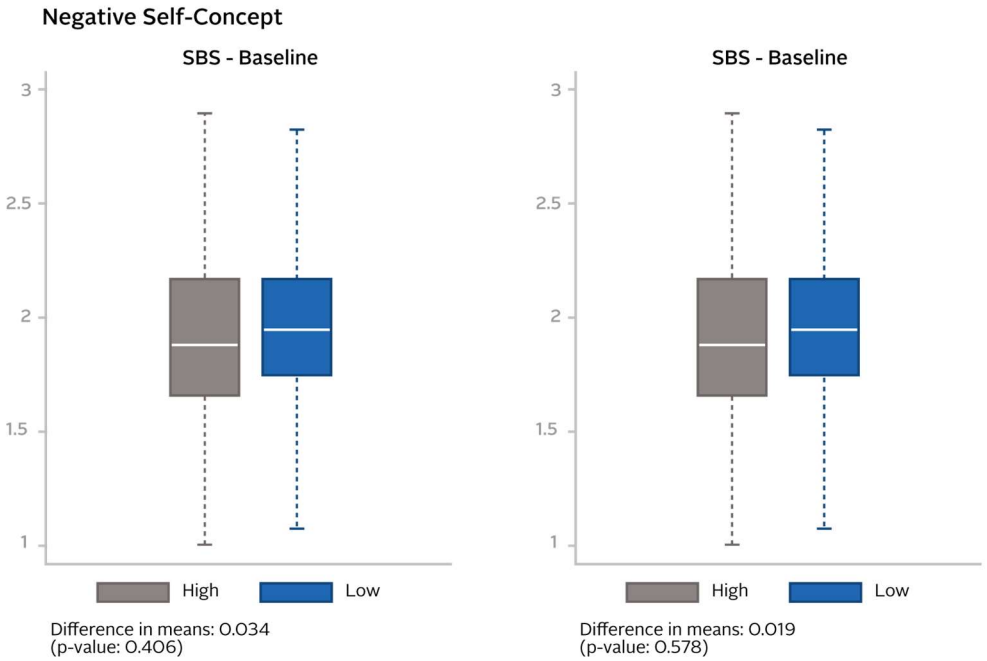


Figure A. 13 Box-plot by SES, Guatemala, Baseline and Endline: HOTS

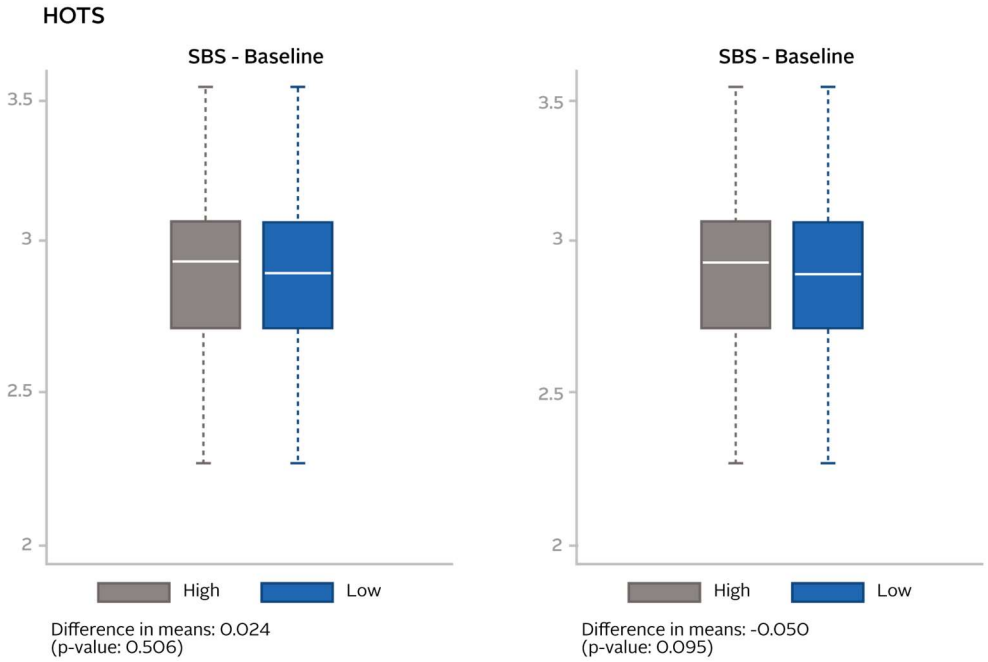


Figure A. 14 Box-plot by SES, Uganda, Baseline and Endline: HOTS

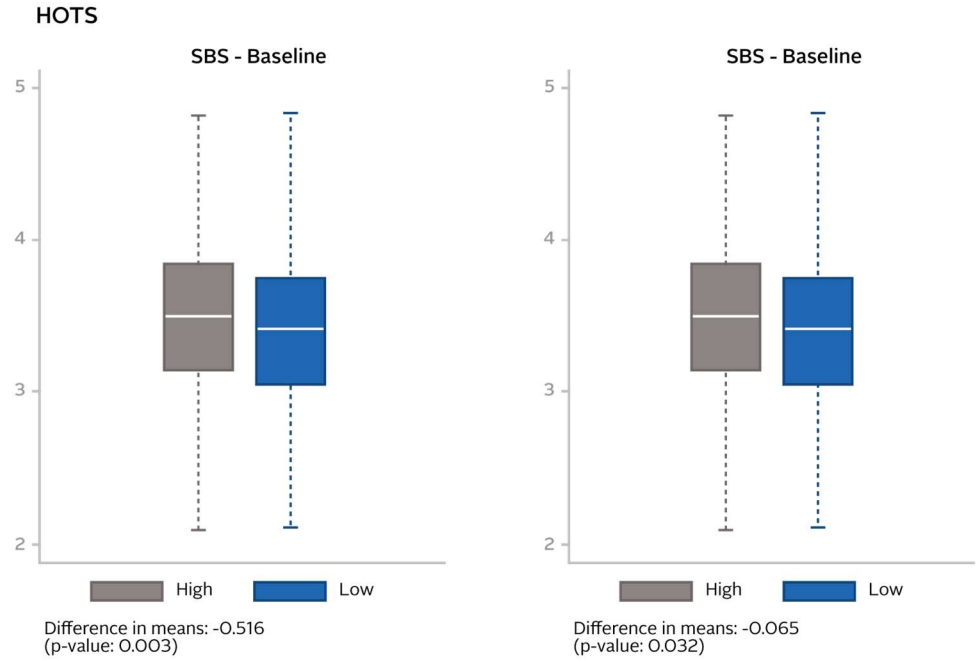


Figure A. 15 Box-plot by SES, Guatemala, Baseline and Endline: Social and Communication Skills

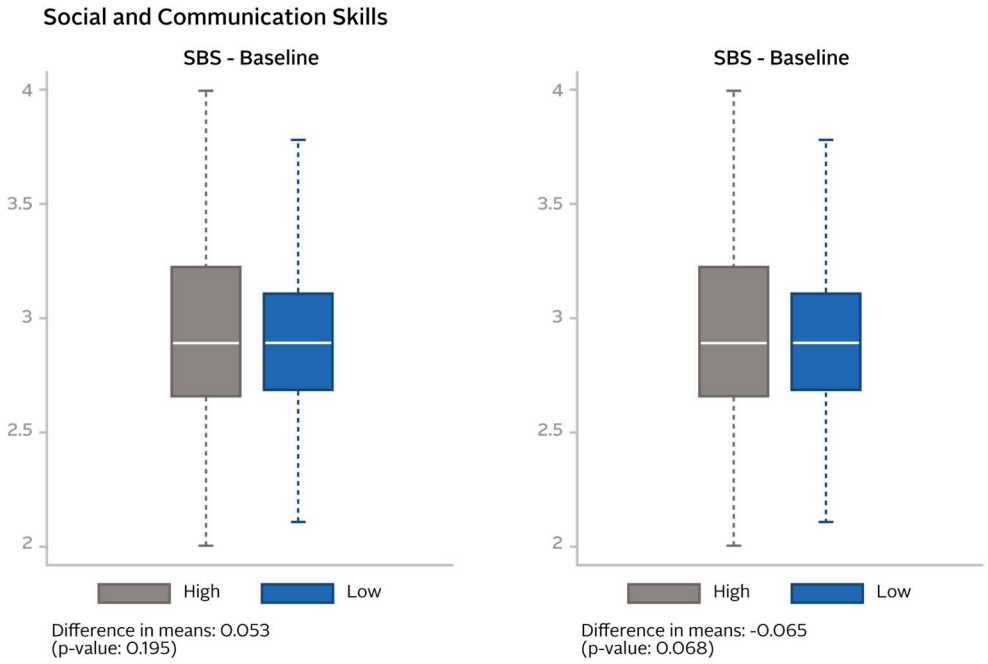
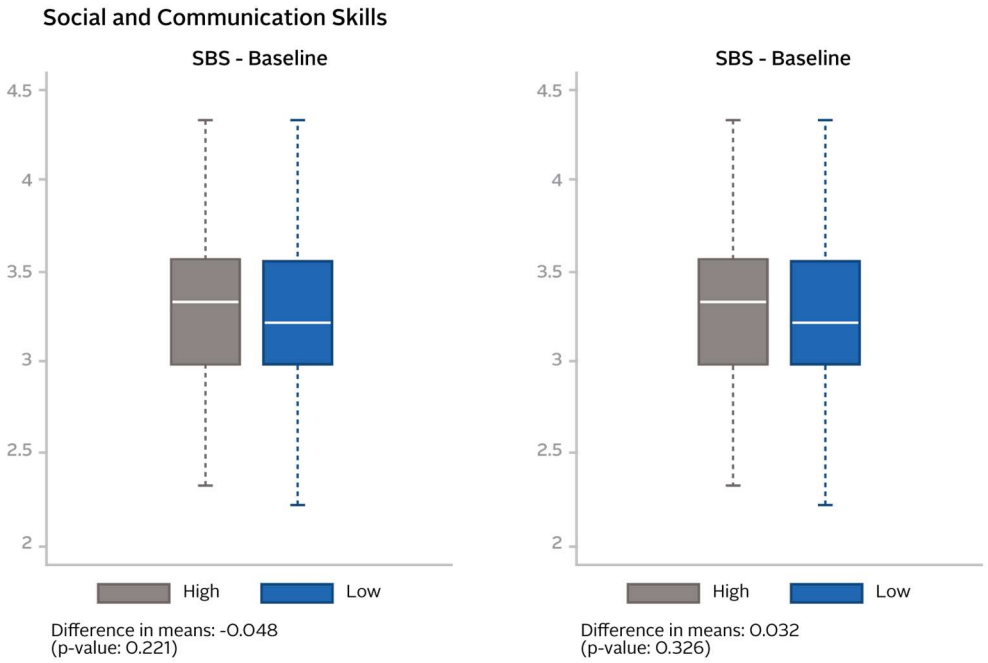


Figure A. 16 Box-plot by SES, Uganda, Baseline and Endline: Social and Communication Skills



The distribution for each scale across rural and urban areas are shown in Figure A.17 – A.20. Youth from rural and urban areas show very similar distribution for the positive self-concept scale, while youth

from the rural area show higher scorings on the negative self-concept scale (difference in means significant only at endline). For the HOTS scale, the distribution shows a large variation for rural youth, with a higher concentration of youth scoring below the median when compared to the urban area (difference in means not significant). Youth from the rural area show higher scoring for the social and communication skills, although the difference in means in only significant at baseline.

Figure A. 17 Box-plot by Rural/Urban, Guatemala, Baseline and Endline: Positive Self-Concept

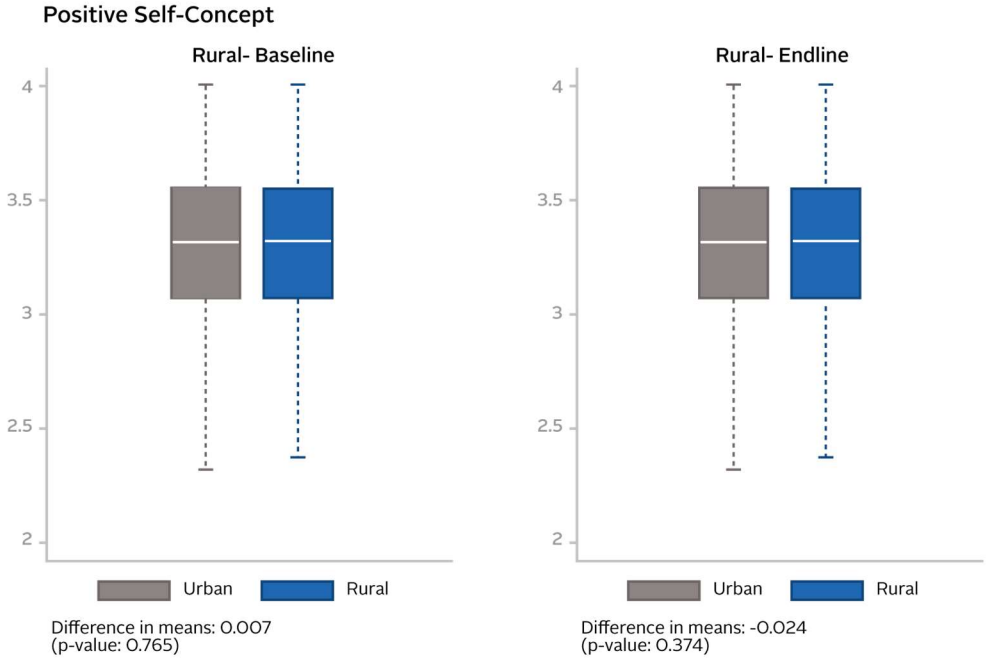


Figure A. 18 Box-plot by Rural/Urban, Guatemala, Baseline and Endline: Negative Self-Concept

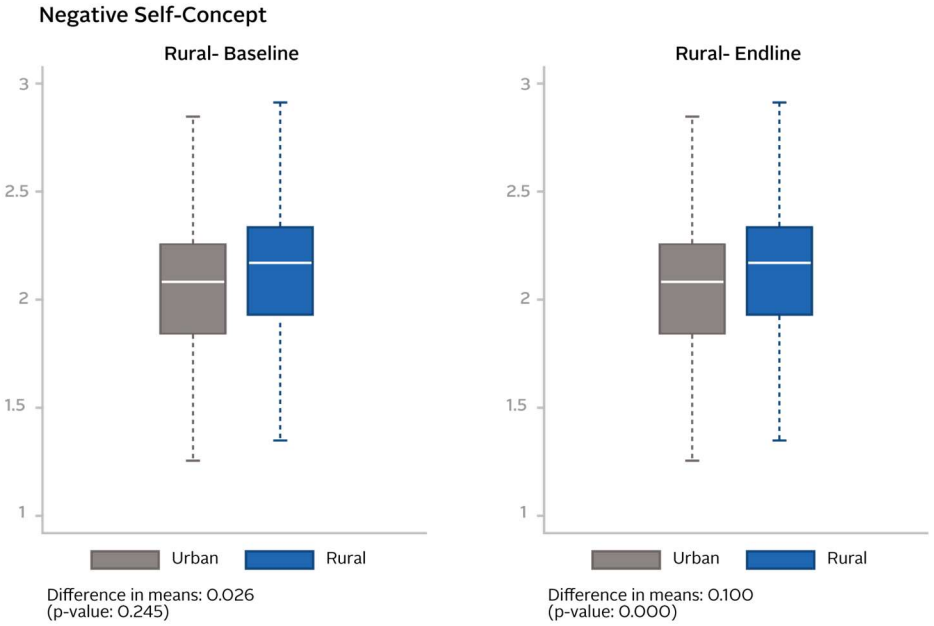


Figure A. 19. Box-plot by Rural/Urban, Guatemala, Baseline and Endline: HOTS

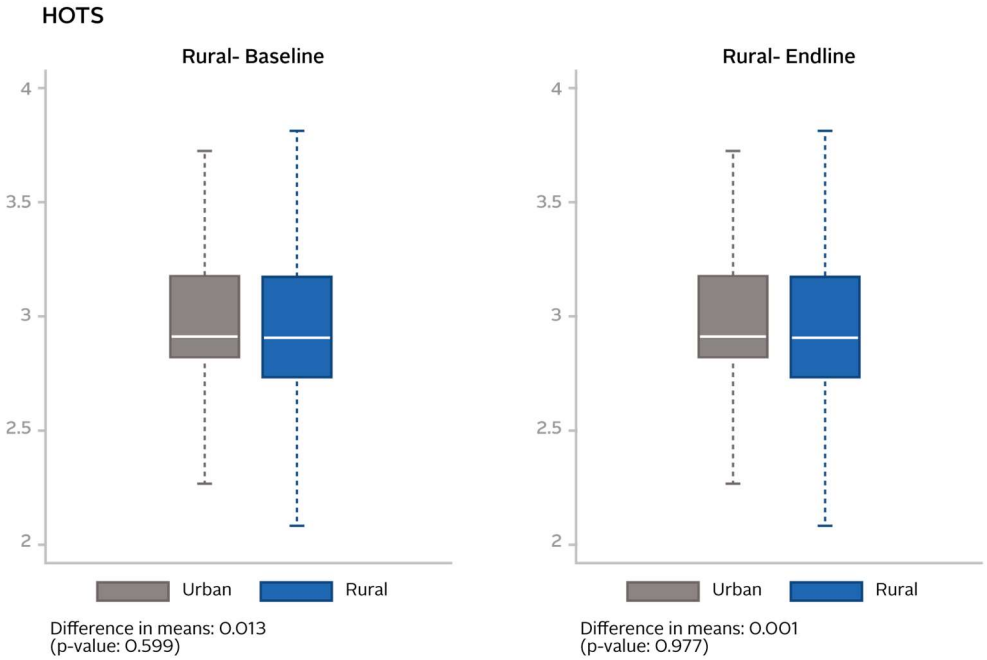
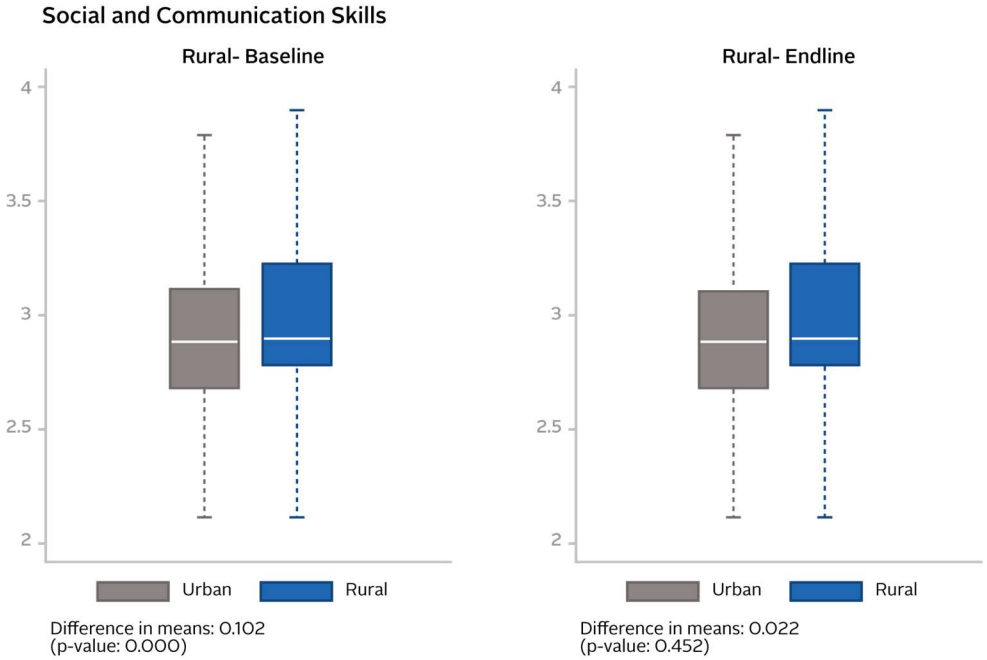


Figure A. 20. Box-plot by Rural/Urban, Guatemala, Baseline and Endline: Social and Communication Skills



Figures A.21 – A.24 show the distribution for each scale by language at home (Spanish and other). Youth that reported speaking Spanish at home show a large variation for the positive self-concept scale (difference in means not significant); lower scoring and variation for the negative self-concept scale

(difference in means significant at both periods); higher scoring for the HOTS scale (difference in means not significant); and larger variation for social and communication skills, with a higher concentration of youth scoring above the median (difference in means significant at endline).

Figure A. 21 Box-plot by Language at Home, Guatemala, Baseline and Endline: Positive Self-Concept

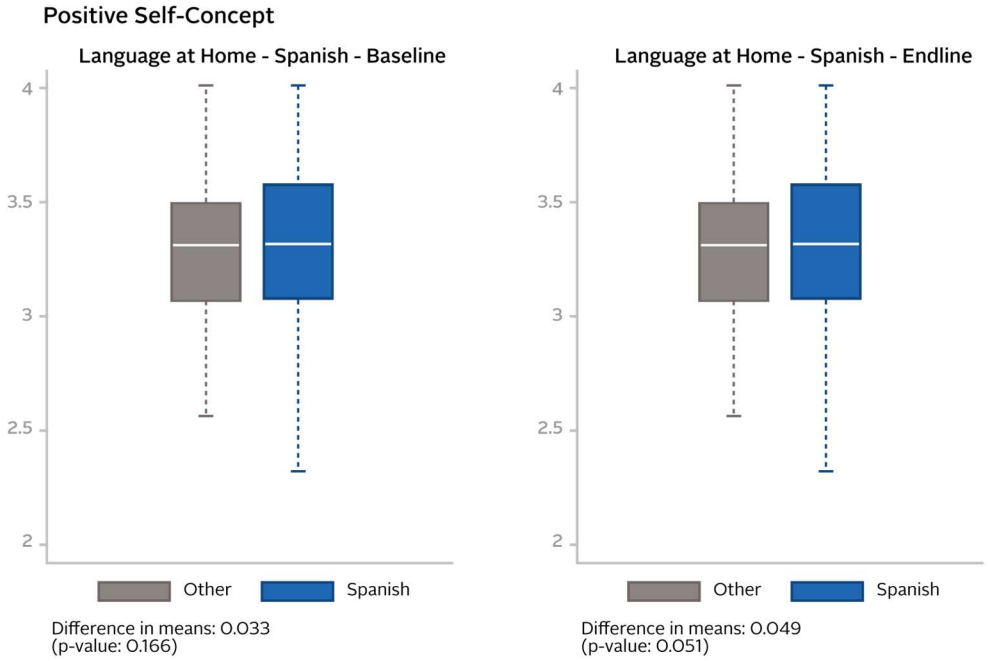


Figure A. 22 Box-plot by Language at Home, Guatemala, Baseline and Endline: Negative Self-Concept

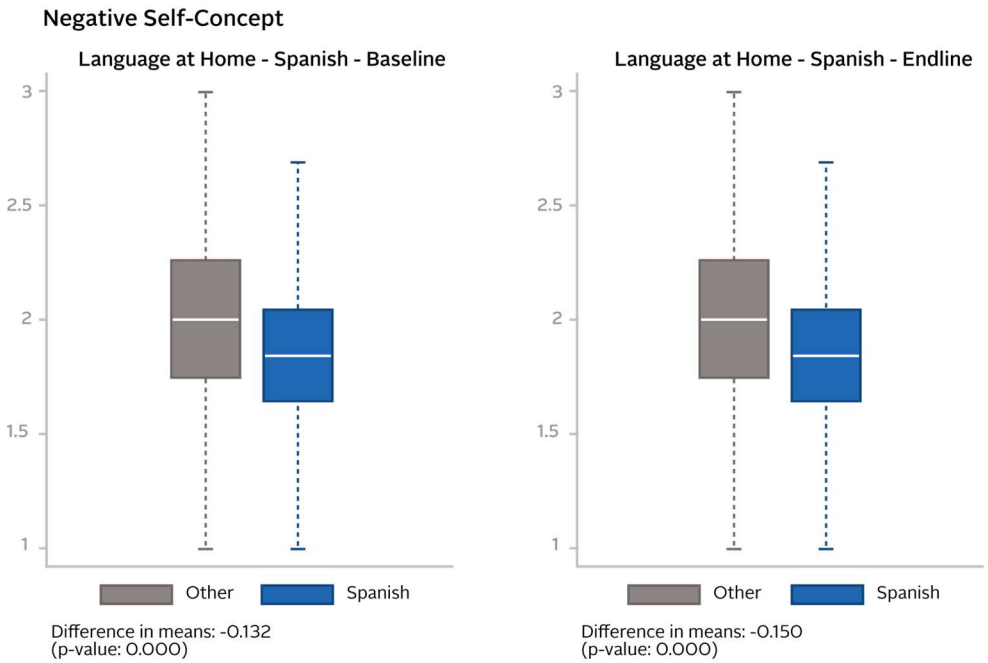


Figure A. 23 Box-plot by Language at Home, Guatemala, Baseline and Endline: HOTS

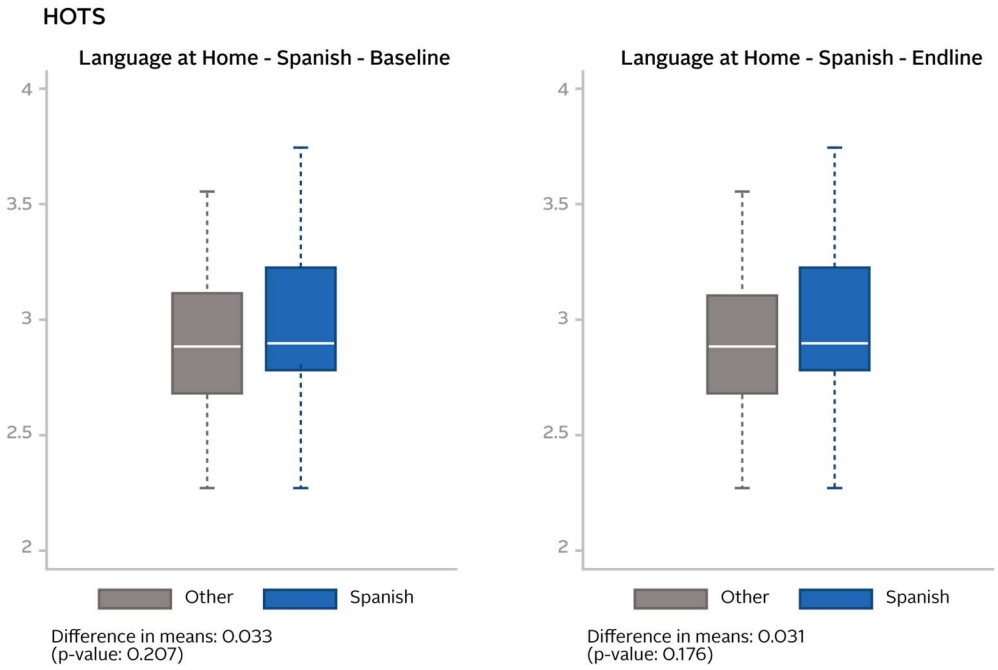
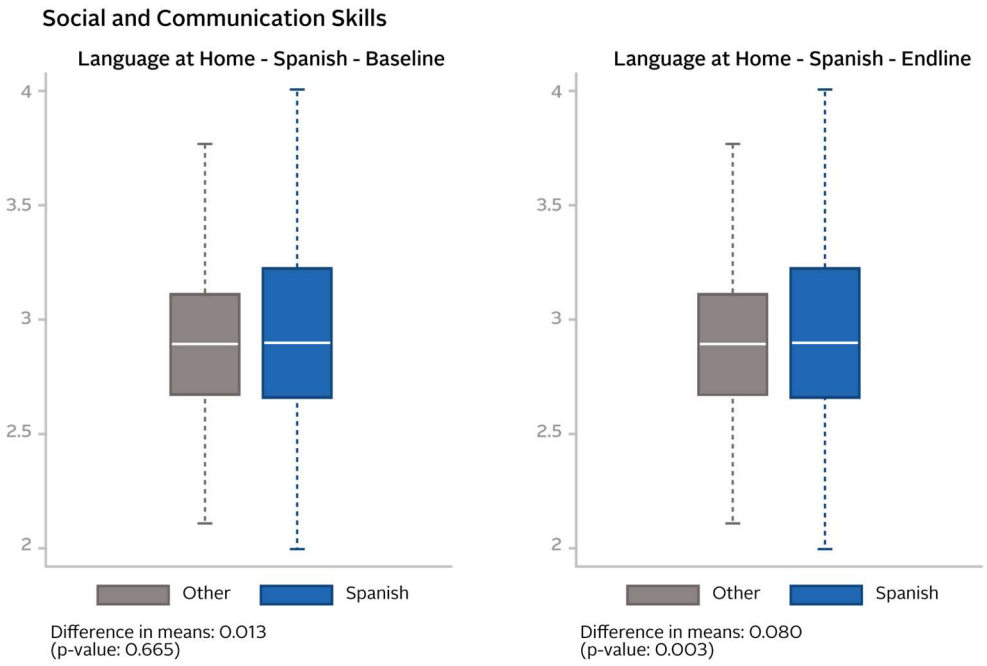


Figure A. 24. Box-plot by Language at Home, Guatemala, Baseline and Endline: Social and Communication Skills



Appendix 3. Additional Data from Exploratory Factor Analysis

Table A. 5 Factor Loadings from Exploratory Factor Analysis, Uganda Baseline

1	2	3	4	Uniqueness	Variable Name	Question
-0.07	-0.02	0.19	0.05	**0.955**	self_efficacy1	18. How often do you believe that: it is hard for you to solve your problems?
-0.04	0.14	**0.29**	-0.11	0.882	self_efficacy2	19. How often do you believe that: there are many things that you do poorly?
0.37	0.00	-0.02	0.09	0.853	self_efficacy3	20. How often do you believe that: you are good at learning something new?
0.37	0.17	0.03	0.02	0.833	self_efficacy4	21. How often do you believe that: you can do most things if you make the effort
0.52	0.07	0.01	-0.13	0.711	self_efficacy5	22. How often do you believe that: you can do something that will help you succe
-0.11	0.10	**0.34**	-0.01	0.860	self_esteem1	24. How often do you feel that you are not good at all?
0.35	0.11	-0.02	-0.03	0.868	self_esteem2	25. How often have you felt that: the people you live with at home value you?
0.36	0.03	-0.06	0.05	0.861	self_esteem5	26. How often do you feel that you are a valued member of your community?
0.35	0.10	-0.01	0.04	0.866	self_esteem4	27. How often have you felt that: you have a number of good qualities?
0.41	0.02	-0.03	0.12	0.821	self_confidence1	29. How often do you feel good about your skills?
-0.15	-0.09	**0.29**	0.00	0.888	self_confidence2	30. How often do you feel not sure that you can be successful?
-0.32	0.06	**0.31**	0.05	0.796	self_confidence3	31. How often do you feel that you don't trust your skills?
0.47	0.15	-0.05	-0.05	0.756	self_confidence4	32. How often do you feel confident in yourself?
-0.08	0.14	**0.36**	0.10	0.838	self_awareness1	34. How often do you find it hard to know how you are feeling?
0.36	0.02	-0.06	0.09	0.859	self_awareness2	35. How often do you know what you are good at?
0.26	0.19	0.03	-0.02	0.898	self_awareness3	36. How often do you know how you are feeling inside at any particular moment?
0.09	0.18	-0.01	0.11	**0.947**	self_awareness4	37. How often do you know how you make other people feel?
0.61	-0.15	-0.03	0.00	0.604	self_belief1	39. How often do you see that your future will be happy?
0.58	-0.02	0.04	0.04	0.661	self_belief2	40. How often do you believe that you will reach your future
0.46	-0.03	-0.03	0.06	0.780	self_belief3	41. How often do you know that you are going to be fine?
0.45	0.10	0.03	-0.01	0.790	self_belief4	42. How often do you believe you can make things happen that will improve your l
0.26	0.12	0.00	0.05	**0.917**	gratification1	44. How often do you save your money for something you want to
0.00	**0.30**	0.20	-0.11	0.859	gratification2	45. How often do you find it challenging to wait for something?
0.04	0.13	0.13	0.08	**0.959**	gratification3	46. How often would you prefer to get one pen now rather than many pens later?
-0.03	-0.04	**0.33**	0.01	0.891	impulses1	48. How often do you do things without thinking about what you're doing?
0.07	0.02	0.17	0.05	**0.963**	impulses2	49. In the past month, how often have you interrupted your agemate when they wer
0.21	**0.32**	-0.09	0.06	0.845	impulses3	50. How often do you think through things before you do them?
0.12	0.16	-0.07	0.14	**0.937**	attention1	52. In the past month, how often have you finished the work that you set out to ?
0.06	-0.08	**0.35**	-0.04	0.864	attention2	53. In the past month, how often have you been unable to pay attention?
0.15	0.09	-0.02	0.09	**0.962**	attention3	54. In the past month, how often have you kept doing something that you should d
0.00	0.15	**0.38**	-0.05	0.832	attention4	55. In the past month, how often have you found it difficult to start your work?
0.11	**0.41**	-0.05	0.01	0.822	emotions1	61. In the past month, how often have you done things to control your anger or t
-0.03	-0.03	**0.31**	0.02	0.900	emotions2	62. In the past month, how often have you been annoyed by little things, like if
0.13	**0.35**	-0.05	-0.07	0.852	emotions3	63. In the past month, how often have you remained calm when a friend tells you
0.14	**0.36**	0.04	-0.03	0.852	regulate_behaviors1	65. In the past month, how often were you able to stop yourself when you were go
0.06	-0.06	**0.48**	-0.07	0.759	regulate_behaviors2	66. In the past month, how often have you refused to follow instructions?
0.27	0.14	-0.02	0.05	**0.903**	regulate_behaviors3	67. In the past month, how often have you got your work done immediately instead
-0.04	-0.02	**0.27**	-0.07	**0.919**	thrill_seeking1	69. How often do you do crazy things, such as drinking alcohol, even if they are
0.02	-0.14	**0.31**	0.05	0.885	thrill_seeking2	70. How often do you do what feels good to you without thinking about its result
-0.11	-0.21	**0.29**	0.18	0.830	thrill_seeking3	71. How often do you do something risky because of peer pressure?
-0.08	**0.35**	-0.04	0.17	0.841	problemsolving1	74. In the past month, how often did you take action to solve the problems?
0.03	0.09	0.03	**0.32**	0.886	problemsolving2	75. In the past month, how often did you ask other people for help with the prob
0.10	**0.35**	0.05	0.12	0.851	problemsolving3	76. In the past month, how often did you try to think of different ways to solve
0.00	**0.32**	0.01	**0.26**	0.831	problemsolving4	77. In the past month, how often did you make a plan to solve the problems?
0.09	0.16	0.02	0.13	**0.948**	critical_thinking1	80. How often did you separate the true and false parts of the story?
0.11	**0.31**	0.07	0.12	0.873	critical_thinking2	81. How often did you question why someone in the story did what they did?
0.11	0.16	0.07	**0.30**	0.866	critical_thinking3	82. How often did you connect pieces of evidence together?
0.13	0.16	-0.03	**0.28**	0.875	decisions1	89. Before making the decisions, how often did you collect a lot of information?
0.01	**0.45**	-0.06	0.02	0.797	decisions2	90. Before making the decisions, how often did you think about
0.09	**0.33**	-0.01	0.07	0.876	decisions3	91. Before making the decisions, how often did you consider different options?
-0.03	**0.31**	-0.01	0.07	0.898	social_skills1	97. How often do you avoid making your agemates look bad?
0.05	**0.35**	-0.01	0.12	0.865	social_skills2	98. How often do you find a way to work things out if two of your friends quarre
0.22	0.13	-0.01	0.15	**0.912**	social_skills3	99. How often do you do your part when you work in a group?
0.05	**0.29**	-0.03	0.17	0.880	social_skills4	100. How often do you relate well with people of different backgrounds?
0.20	-0.07	0.01	**0.34**	0.840	social_skills5	101. How often do you find it easy to make friends?
-0.04	**0.46**	-0.20	-0.01	0.749	social_skills6	102. How often do you control your anger when you have a misunderstanding with a
0.18	0.10	0.02	0.11	**0.944**	social_skills7	103. How often do you respect views that differ from your own?
0.04	0.05	-0.06	**0.32**	0.891	communication1	109. How often do you write a story or letter well?
0.07	0.14	-0.06	0.25	**0.911**	communication2	110. How often do you listen to your agemates' ideas?
-0.02	0.16	-0.11	**0.27**	0.893	communication3	111. How often can you discuss a problem with a friend without making things wor
-0.01	-0.10	**0.31**	-0.08	0.891	communication4	112. How often are you uncomfortable to ask questions in class?
0.03	-0.14	**0.32**	-0.05	0.880	communication5	113. How often are you rude to others?
-0.04	-0.03	-0.04	**0.44**	0.806	communication6	114. How often do you tell others how you feel?

Table A. 6 Factor Loadings from Exploratory Factor Analysis, Guatemala Baseline

1	2	3	4	Uniqueness		
-0.0039	**0.2655**	-0.009	0.1935	0.89	self_efficacy2	4. There are many things that I can't do very well
0.4175	0.1266	0.033	0.1388	0.79	self_efficacy3	5. I'm good at learning new things
0.4016	0.1425	0.1262	0.0926	0.79	self_efficacy4	6. I can do most things if I make an effort
0.516	0.1463	0.0438	0.0339	0.71	self_efficacy5	7. I can do things that will help me succeed in life
0.1982	**0.4717**	-0.0477	0.0736	0.73	self_esteem1	8. I think I am no good at all.
0.451	0.1553	0.0691	0.024	0.77	self_esteem2	9. I feel valued by the people I live with at home
0.3209	0.0384	0.1068	0.2486	0.82	self_esteem3	10. I'm a valued member of my community
0.4309	0.1608	-0.0156	0.1631	0.76	self_esteem4	11. I have a number of good qualities.
0.5818	0.0858	0.0871	0.0718	0.64	self_esteem5	12. I like myself just the way I am
0.5706	0.0895	0.1129	0.0903	0.65	self_confidence1	13. I feel good about my skills
0.1909	**0.3559**	-0.0203	0.039	0.83	self_confidence2	14. I'm not sure I can be successful
0.2723	**0.4841**	-0.0175	-0.0105	0.69	self_confidence3	15. I'm not confident about my skills
0.5638	0.0761	0.1566	0.0433	0.65	self_confidence4	16. I feel confident in myself
-0.0745	**0.2776**	0.0145	0.1095	**0.91**	self_awareness1	17. It is hard to know what I'm feeling
0.4594	0.1451	0.1479	0.1662	0.72	self_awareness2	18. I know what I'm good at
0.2951	0.0263	0.1609	0.2168	0.84	self_awareness3	19. I know how I'm feeling inside at any particular moment
0.5303	0.0116	0.1379	0.1483	0.68	self_belief1	20. My future will be happy
0.5467	0.1058	0.1258	0.0443	0.67	self_belief2	21. I can achieve most of my future goals.
0.4684	-0.0177	-0.0039	0.2011	0.74	self_belief3	22. I know I'm going to be fine
0.5202	0.0795	0.1768	0.0823	0.69	self_belief4	23. I can make things happen that will improve my life
0.1161	**0.49**	0.1844	0.0089	0.71	impulses1	24. I do things before I think through them
0.2898	0.2066	**0.3588**	0.1264	0.73	impulses2	25. I think carefully before doing anything
0.0022	**0.3372**	-0.0162	0.2366	0.83	attention1	26. I have a hard time concentrating on one thing.
0.0112	**0.449**	0.1319	0.0785	0.77	attention2	27. I have difficulty starting tasks
0.1121	0.1002	**0.3207**	**0.3097**	0.78	emotions1	30. When things go wrong for me, I'm good controlling my temper
0.0791	**0.4039**	-0.0559	0.0179	0.83	emotions2	31. I'm easily annoyed by little things (like if someone steps on my shoe)
0.0475	0.0718	0.251	0.0728	**0.92**	emotions3	32. If a friend tells me I did something wrong, I can stay calm
0.2747	0.2373	0.1792	0.0422	0.83	regulate_behaviors	33. If I'm doing something that I know I would regret, I'm able to stop before it
0.2722	0.2324	0.2235	**0.3044**	0.73	regulate_behaviors	34. I'm good at following instructions
0.1254	**0.5152**	0.1569	0.0523	0.69	thrill_seeking1	35. I do whatever feels good to me, without thinking about the results
0.1933	**0.3679**	0.057	-0.091	0.82	thrill_seeking2	36. If my friends are doing something risky, I will do it with them
0.1367	0.1156	**0.4672**	0.1348	0.73	problemsolving1	37. I took action to solve the problems
0.1364	0.063	**0.3655**	0.052	0.84	problemsolving2	38. I asked other people for help to solve the problems
0.2095	0.0717	**0.5063**	0.1415	0.67	problemsolving3	39. I tried to think of different ways to solve the problems
0.1416	0.0078	**0.4473**	0.1718	0.75	problemsolving4	40. I made a plan to solve the problems
0.1209	0.0736	**0.2996**	-0.1037	0.88	critical_thinking1	41. I questioned why someone in the story did what they did
0.1276	-0.0475	**0.2648**	0.056	**0.91**	critical_thinking2	42. I connected pieces of evidence together
0.1875	0.0382	**0.3964**	0.1565	0.78	decisions1	45. I collected a lot of information before making the decision
0.0955	0.0548	**0.3593**	0.0598	0.86	decisions2	46. I thought about how other people would be affected before making the
0.1696	-0.0243	**0.425**	0.1026	0.78	decisions3	47. I considered different options before making the decision
0.2172	0.1702	0.2469	**0.3113**	0.77	social_skills1	50. I'm able to stand up for myself without putting others down
0.194	0.1757	0.1553	0.2199	0.86	social_skills2	51. I find a way of working things out if two of my friends quarrel
0.2453	0.1084	0.1476	**0.406**	0.74	social_skills3	52. I get along well with people from different backgrounds
0.1856	0.0135	0.1239	**0.4616**	0.74	social_skills4	53. I find it easy to make friends
0.1237	0.1243	**0.2897**	**0.3211**	0.78	social_skills5	54. I can control my anger when I have a misunderstanding with a friend
0.2311	0.0695	0.1278	**0.3073**	0.83	communication1	57. I write well.
0.1845	0.0623	0.2581	**0.461**	0.68	communication2	58. I am good at resolving disagreements.
0.1478	0.0353	0.0867	**0.4069**	0.80	communication3	59. It is easy for me to ask questions in a public setting.
0.1325	**0.4082**	0.063	0.1417	0.79	communication4	60. I am rude to others.
0.0871	-0.0345	0.0668	**0.4001**	0.83	communication5	61. It is easy for me to share my feelings with others.

Appendix 4. Background on Measurement Invariance Analysis Process

We use multiple group confirmatory factor analysis (MG-CFA) (Meredith, 1993) to test for MI. MG-CFA can be seen as an extension of CFA, since it tell us about model fit while also telling us whether the model fit is different for the different groups. MG-CFA allows us to progressively impose constraints (by assuming that the factor loadings at the same for the groups) on the parameter estimates (specifically on the factor loadings and on the item intercepts) and then assess the level of equivalency between the groups by looking at the overall model fit (Vandenberg & Lance, 2000).

MG-CFA consists of several steps, with each step posing a more difficult "test" to the model. Our first step in MG-CFA was to fit a model in which factor loadings and item intercepts may vary across the groups, testing whether the overall factor structure holds up similarly across the groups. This finding is known as *configural invariance*. Next, we hold constant the factor loadings across the groups to test whether the underlying factors are on the same metric. This finding would indicate *metric invariance*. Finally, we hold constant both the factor loadings and item intercepts across groups, testing whether the latent scores are on the same metric *and* have the same origin. This is known as *scalar invariance*.

As increasingly restrictive models are fit (i.e., from configural to metric to scalar) the overall quality of the fit deteriorates. We determine the level of measurement non-invariance by comparing the fit statistics of these models. For each model, we report the Chi-square fit statistic, the degrees of freedom, the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI) (). The RMSEA and the SRMR are *absolute* measures of fit (i.e., values closer to zero indicate a better fit), where values below 0.05 for the RMSEA and of 0.08 for the SRMR are commonly accepted as representing a good fit. Unlike Chi-square, these measures are not sensitive to sample size, and therefore, are useful for gauging overall levels of invariance. The CFI and TLI are *relative* fit indexes (i.e., the model fit is placed on the continuum from the null model to the perfect model; values closer to 1 indicate a better fit), where values above 0.95 are commonly accepted as representing a good fit—which, while useful, are redundant in situations where measures of absolute fit are within an acceptable range.

We use Chen (2007) as a reference in determining the level of invariance. For metric invariance, a change larger than 0.030 in the SRMR or larger than 0.015 in the RMSEA is indicative of non-invariance. In turn, for testing scalar invariance, a change larger than 0.010 in SRMR or larger than 0.015 in RMSEA would indicate non-invariance. All the configural-to-metric and metric-to-scalar changes in the model fit statistics are below these thresholds, which means the assessment demonstrates metric and scalar invariance for the constructs across all characteristics (see). **This indicates that the overall scale structure fits well in applications across contexts such as Uganda and Guatemala, and across the subgroups within these contexts.**

Table A. 7 Measurement Invariance Analysis, Fit Statistics

	CHISQ	DF	SRMR	RMSEA	RMSEA.L	RMSEA.U	CFI	TLI
Guatemala x Uganda endline								
Model 1: Configural	5269.4	2148	0.053	0.04	0.039	0.042	0.968	0.966
Model 2: Metric Invariance (loadings)	6693.9	2192	0.061	0.048	0.047	0.049	0.953	0.952
Model 3: Scalar Invariance (intercepts)	6943.5	2284	0.055	0.048	0.047	0.049	0.951	0.952

Guatemala baseline x Guatemala endline								
Model 1: Configural	4617.3	2148	0.047	0.038	0.037	0.04	0.827	0.818
Model 2: Metric Invariance (loadings)	4686.1	2192	0.049	0.038	0.037	0.04	0.825	0.82
Model 3: Scalar Invariance (intercepts)	4877.2	2236	0.05	0.039	0.037	0.04	0.815	0.813
Guatemala baseline: gender								
Model 1: Configural	3719	2148	0.057	0.044	0.041	0.046	0.763	0.752
Model 2: Metric Invariance (loadings)	3767.5	2192	0.058	0.043	0.041	0.046	0.763	0.756
Model 3: Scalar Invariance (intercepts)	3858.9	2236	0.059	0.044	0.041	0.046	0.756	0.753
Guatemala baseline: SES								
Model 1: Configural	3719	2148	0.057	0.044	0.041	0.046	0.763	0.752
Model 2: Metric Invariance (loadings)	3767.5	2192	0.058	0.043	0.041	0.046	0.763	0.756
Model 3: Scalar Invariance (intercepts)	3858.9	2236	0.059	0.044	0.041	0.046	0.756	0.753
Guatemala endline: gender								
Model 1: Configural	3839.4	2148	0.058	0.045	0.043	0.048	0.792	0.782
Model 2: Metric Invariance (loadings)	3894.1	2192	0.061	0.045	0.043	0.047	0.791	0.785
Model 3: Scalar Invariance (intercepts)	3964	2236	0.062	0.045	0.043	0.047	0.788	0.786
Guatemala endline: SES								
Model 1: Configural	3839.4	2148	0.058	0.045	0.043	0.048	0.792	0.782
Model 2: Metric Invariance (loadings)	3894.1	2192	0.061	0.045	0.043	0.047	0.791	0.785
Model 3: Scalar Invariance (intercepts)	3964	2236	0.062	0.045	0.043	0.047	0.788	0.786
Uganda endline: gender								
Model 1: Configural	3394.7	2148	0.047	0.034	0.032	0.037	0.83	0.821
Model 2: Metric Invariance (loadings)	3433.7	2192	0.049	0.034	0.032	0.036	0.83	0.825
Model 3: Scalar Invariance (intercepts)	3529.3	2236	0.05	0.034	0.032	0.036	0.823	0.822
Uganda endline: SES								
Model 1: Configural	3394.7	2148	0.047	0.034	0.032	0.037	0.83	0.821
Model 2: Metric Invariance (loadings)	3433.7	2192	0.049	0.034	0.032	0.036	0.83	0.825
Model 3: Scalar Invariance (intercepts)	3529.3	2236	0.05	0.034	0.032	0.036	0.823	0.822

Note: delta SRMR rule = 0.03; delta RMSE rule = 0.015

Appendix 5. Background on DIF Analysis

The DIF analysis was carried out via a model-based framework with the ordinal response scale as the outcome and group membership and a measure of the latent trait as the predictors (Agresti, 1990). For each item, three proportional-odds logistic models were fit: 1) latent trait only; 2) latent trait with an indicator for one of the groups; and 3) interaction between the latent trait and the group indicator. A graded response model was used to obtain the estimate of the latent trait (De Boeck and Wilson, 2004; Choi, Gibbons and Crane, 2011).

In practice, all items exhibit some degree of differential functioning. Further, there is considerable disagreement in the criteria to use in comparing the model likelihoods and determining if an item has a relevant differential functioning. For this reason, we used three different criteria (a likelihood ratio Chi-square test, an R squared test and a Beta test) to determine which items, if any, were consistently marked as having DIF, and then explored if the DIF in those items resulted in relevant differences at the construct level. As noted in the narrative of this report, despite identifying DIF in some items, we did not observe a magnitude that would lead to noticeable differences at the construct level, reaffirming the stability of the tool in cross-cultural use.

Table A. 8 Number of items flagged for DIF by factor and comparison, using three criteria: (1) Chi square at 0.01 – (2) R2 – (3) Beta

Comparison	Factor			
	(1) Positive Self Concept 16 items	(2) Negative Self Concept 12 items	(3) HOTS 11 items	(4) Social and Communication Skills 9 items
By gender				
Guatemala Baseline		thrill_seeking1 communication4		social_skills1 social_skills2 communication5
Guatemala Endline	self_efficacy3 self_confidence4	thrill_seeking1 communication4		social_skills5
Uganda Endline				communication1
By SES (low vs high)				
Guatemala Baseline			critical_thinking1 critical_thinking2	social_skills4 communication5
Guatemala Endline				
Uganda Endline		emotions2	decisions2	

Table A. 9 Items flagged for DIF for gender and SES comparisons

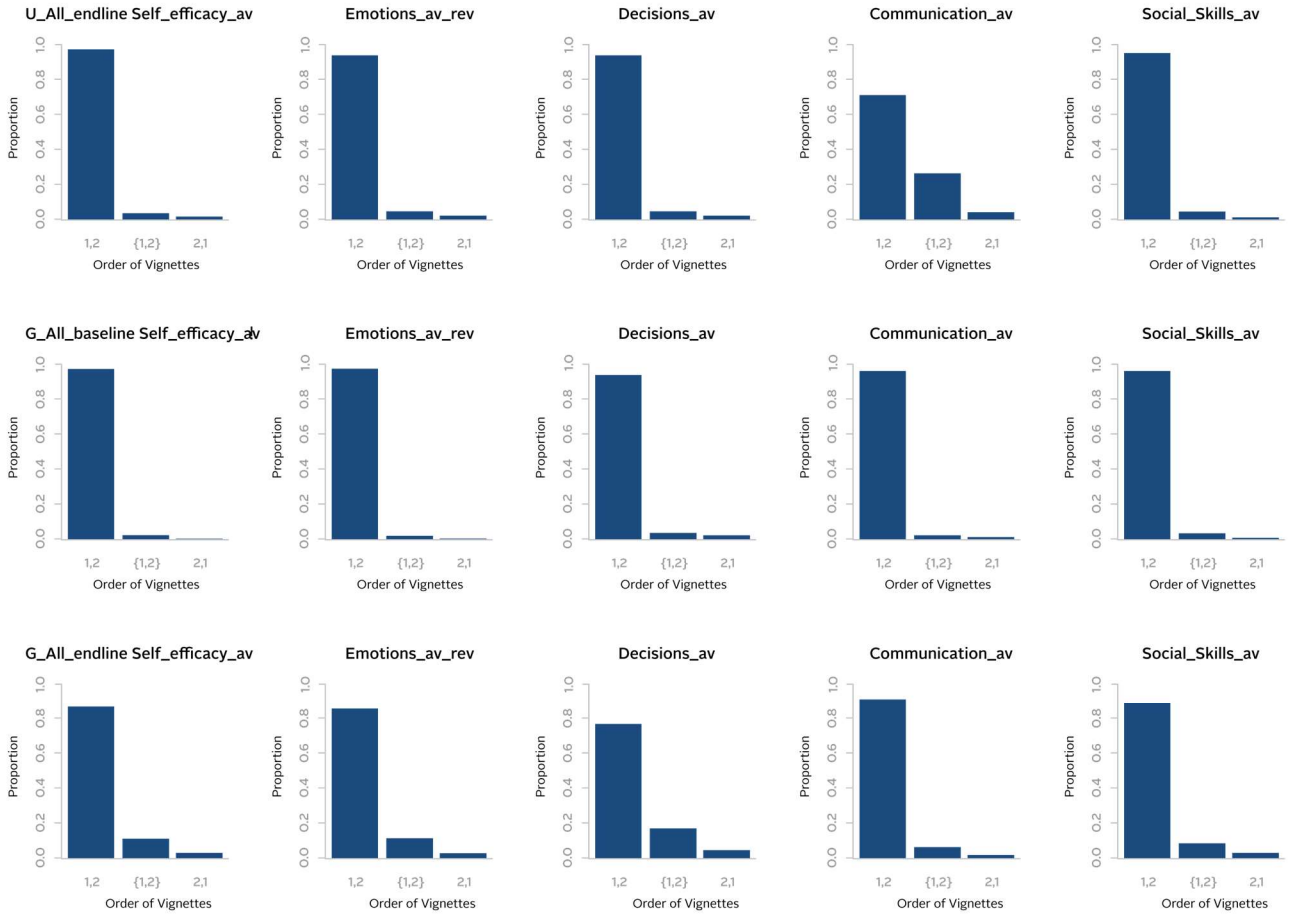
Comparison	Factor			
	(1) Positive Self Concept 16 items	(2) Negative Self Concept 12 items	(3) HOTS 11 items	(4) Social and Communication Skills 9 items
By gender				
Guatemala Baseline		thrill_seeking1 communication4		social_skills1 social_skills2 communication5
Guatemala Endline	self_efficacy3 self_confidence4	thrill_seeking1 communication4		social_skills5
Uganda Endline				communication1
By SES (low vs high)				
Guatemala Baseline			critical_thinking1 critical_thinking2	social_skills4 communication5
Guatemala Endline				
Uganda Endline		emotions2	decisions2	

Appendix 6. Additional Data from Anchoring Vignette Analysis

Table A. 9 Scale means for raw and anchoring vignette (AV) adjusted scores

		Positive self-concept		Negative self-concept		HOTS		Social & communication skills		
	N	Raw	AV	Raw	AV	Raw	AV	Raw	AV 1	AV 2
Uganda Endline	1010	3.48	3.69	2.03	2.55	3.24	3.51	3.23	4.32	3.44
Guatemala Baseline	794	3.38	3.86	2.04	2.52	2.98	3.55	2.97	3.52	3.38
Guatemala Endline	784	3.27	3.66	2.17	2.69	2.94	3.39	2.93	3.40	3.21
Uganda Retest 1	57	3.47	3.61	1.97	2.63	3.27	3.50	3.24	4.35	3.39
Uganda Retest 2	57	3.49	3.66	1.84	2.49	3.30	3.47	3.30	4.47	3.41
Guatemala Retest 1	126	3.33	3.64	2.09	2.64	2.98	3.40	2.97	3.56	3.27
Guatemala Retest 2	126	3.29	3.61	2.05	2.65	3.00	3.38	3.03	3.52	3.43

Figure A. 25 Anchoring Vignettes, Response Analysis



Appendix 7. Additional Information on Program Staff Tool Analysis

Table A. 10 Mapping of Program Staff Items to Youth Items and Factors

Program Staff Item	Youth Item: Version 1	Youth Item: Version 2	Youth Factor	Program Staff Subscale
How often does the youth choose to actively participate in the program to build their skills for the future instead of doing something else they would rather do at that moment?	How often do you want to save money for something you want to buy later?	No equivalent	N/A	N/A
	How often do you find it challenging to wait for something?	No equivalent	N/A	N/A
	How often would you prefer to get one pen now rather than many pens later?	No equivalent	N/A	N/A
How often does the youth think through things before doing them?	How often do you do things without thinking about what you're doing?	I do things without thinking about them.	Factor 3	Subscale 3 (v2, v4)
	In the past month, how often have you interrupted your friend when they were telling a story?	No equivalent	N/A	N/A
	How often do you think through things before you do them?	I think carefully before doing anything.	Factor 2	Subscale 2 (v1, v3)
How often is the youth unable to pay attention?	In the past month, how often have you been unable to pay attention?	I have a hard time concentrating on one thing.	Factor 3	Subscale 3 (v1, v3)
How often does the youth complete project activities through to the end?	In the past month, how often have you finished the work that you set out to do despite challenges?	No equivalent	N/A	N/A
None	In the past month, how often have you kept doing something that you should do even if you didn't like it, such as homework?	No equivalent	N/A	N/A
How often does the youth have difficulty starting tasks?	In the past month, how often have you found it difficult to start your work?	I have difficulty starting tasks	Factor 3	Subscale 3 (v1, v3)

How often does the youth manage to stay calm?	In the past month, how often have you done things to control you anger or temper, for example when you have quarreled with your friend?	When things go wrong for me, I'm good at controlling my temper.	Factor 2	Subscale 2 (v1, v3)
	In the past month, how often have you been annoyed by little things, like if someone steps on your shoe?	I'm easily annoyed by little things (like if someone steps on my shoe).	Factor 3	Subscale 3 (v2, v4)
	In the past month, how often have you remained calm when a friend told you that you did something poorly?	If a friend tells me I did something wrong, I can stay calm.	Dropped due to low loading	N/A
How often does the youth follow instructions?	In the past month, how often have you refused to follow instructions?	I'm good at following instructions.	Factor 4	Subscale 4 (all versions)
How often is the youth ready to actively participate in the program?	No equivalent	No equivalent	N/A	N/A
How often does the youth meet deadlines?	In the past month, how often have you got your work done immediately instead of waiting until the last minute?	No equivalent	N/A	N/A
How often does the youth engage in risky activities?	How often do you do crazy things, such as drinking alcohol, even if they are a little dangerous?	No equivalent	N/A	N/A

	How often do you do what feels good to you without about its results?	I do whatever feels good to me, without thinking about the results.	Factor 3	Subscale 3 (all versions)
	How often do you do something risky because of peer pressure?	If my friends are doing something risky, I will do it with them.	Factor 3	Subscale 3 (all versions)
How often does the youth to think of different ways to solve the problem?	In the past month, how often did you try to think of different ways to solve the problems?	I tried to think of different ways to solve the problem.	Factor 2	Subscale 2 (all versions)
	In the past month, how often did you take action to solve the problems?	I took actions to solve the problems.	Factor 2	Subscale 2 (all versions)
	In the past month, how often did you ask other people for help with the problems?	I asked other people for help to solve the problems.	Factor 2	Subscale 2 (all versions)
	In the past month, how often did you make a plan to solve the problems?	I made a plan to solve the problems.	Factor 2	Subscale 2 (all versions)
How often does the youth exercise strong reasoning and critical thinking?	How often did you separate the true and false parts of the story?	No equivalent	N/A	N/A
	How often did you question why someone in the story did what they did?	I questioned why someone in the story did what they did.	Factor 2	Subscale 2 (all versions)
	How often did you connect pieces of evidence together?	I connected pieces of the story.	Factor 2	Subscale 2 (all versions)

How often does the youth use knowledge and information before making decisions?	Before making the decisions, how often did you collect a lot of information?	I collect a lot of information before making a decision.	Factor 2	Subscale 2 (all versions)
	Before making the decisions, how often did you think about how others would be affected?	I think about how other people would be affected before making a decision.	Factor 2	Subscale 2 (all versions)
	Before making the decisions, how often did you consider different options?	I consider different options before making a decision.	Factor 2	Subscale 2 (all versions)
How often does the youth stand up for himself/herself without making others feel bad?	How often do you avoid making your friends look bad?	I'm able to stand up for myself without making others feel bad.	Factor 4	Subscale 4 (all versions)
How often does the youth find a way to work things out if two of his/her friends have a quarrel?	How often do you find a way to work things out if two of your friends quarrel?	I find a way of working things out if two of my friends quarrel.	Dropped due to low loading	N/A
How often does the youth get along well with people of different backgrounds?	How often do you relate well with people of different backgrounds?	I get along well with people from different backgrounds.	Factor 4	Subscale 4 (all versions)
How often is the youth able to control his/her anger when having a misunderstanding with a friend?	How often do you respect views that differ from your own?	no equivalent	N/A	N/A
	How often do you do your part when you work in a group?	no equivalent	N/A	N/A
	How often do you find it easy to make friends?	I find it easy to make friends	Factor 4	Subscale 4 (all versions)

	How often do you control your anger when you have a misunderstanding with a friend?	I can control my anger when I have a misunderstanding with a friend	Factor 4	Subscale (all versions)
How often does the youth write well?	How often do you write a story or a letter well?	I write well. For example, I write stories or articles or letters or other things like those well.	Factor 4	Subscale 4 (all versions)
How often does the youth speak articulately?	N/A	N/A	N/A	N/A
How often does the youth ask questions in public?	How often are you uncomfortable to ask questions in public?	It is easy for me to ask questions in public.	Factor 4	Subscale 4 (all versions)
How often does the youth make eye contact when talking?	N/A	N/A	N/A	N/A
How often is the youth rude to others?	How often are you rude to others?	I am rude to others.	Factor 3	Subscale 2 (v1, v3)
	How often do you listen to your friends' ideas?	No equivalent	N/A	N/A
	How often can you discuss a problem with a friend without making things worse?	I am god at resolving disagreements.	Factor 4	Subscale 4 (v4)
	How often do you tell others how you feel?	It is easy for me to tell others how I feel.	Factor 4	Subscale 4 (v4)

Table A. 12 Program Staff – Youth Correlation Matrix, with Cronbach's alphas, Uganda Baseline[13]

			Youth			Staff Version 1			Staff Version 2			Staff Version 3			Staff Version 4		
			Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4
			f2_y	f3_y	f4_y	f2_s1	f3_s1	f4_s1	f2_s2	f3_s2	f4_s2	f2_s3	f3_s3	f4_s3	f2_s4	f3_s4	f4_s4
Youth	Factor 2	f2_y	**0.749**														
	Factor 3	f3_y	-0.254	**0.626**													
	Factor 4	f4_y	0.397	-0.086	**0.450**												
Staff Version 1	Factor 2	f2_y	-0.079	-0.045	-0.134	**0.737**											
	Factor 3	f3_y	0.006	-0.053	-0.088	0.568	**0.639**										
	Factor 4	f4_y	-0.072	-0.065	-0.115	0.536	0.448	**0.421**									
Staff Version 2	Factor 2	f2_y	-0.079	-0.066	-0.095	0.839	0.552	0.602	**0.595**								
	Factor 3	f3_y	0.004	-0.052	-0.092	0.639	0.912	0.486	0.527	**0.608**							
	Factor 4	f4_y	-0.068	-0.032	-0.135	0.818	0.700	0.476	0.579	0.622	**0.616**						
Staff Version 3	Factor 2	f2_y	-0.093	-0.042	-0.129	0.969	0.628	0.634	0.905	0.661	0.811	**0.794**					
	Factor 3	f3_y	0.022	-0.062	-0.082	0.500	0.932	0.424	0.440	0.953	0.563	0.529	**0.546**				
	Factor 4	f4_y	-0.055	-0.091	-0.124	0.406	0.580	0.638	0.339	0.445	0.601	0.413	0.428	**0.127**			
Staff Version 4	Factor 2	f2_y	-0.034	-0.086	-0.078	0.776	0.329	0.324	0.848	0.358	0.415	0.719	0.289	0.248	**0.446**		
	Factor 3	f3_y	-0.006	-0.047	-0.096	0.665	0.974	0.491	0.604	0.950	0.727	0.717	0.903	0.575	0.377	**0.681**	
	Factor 4	f4_y	-0.009	-0.042	-0.150	0.833	0.596	0.826	0.694	0.621	0.845	0.877	0.538	0.570	0.440	0.650	**0.719**

[13] Statistics in bold reflect Cronbach's alphas

Table A. 11 Program Staff – Youth Correlation Matrix, with Cronbach's alphas, Uganda Endline

			Youth			Staff Version 1			Staff Version 2			Staff Version 3			Staff Version 4		
			Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4
			f2_y	f3_y	f4_y	f2_s1	f3_s1	f4_s1	f2_s2	f3_s2	f4_s2	f2_s3	f3_s3	f4_s3	f2_s4	f3_s4	f4_s4
Youth	Factor 2	f2_y	**0.667**														
	Factor 3	f3_y	-0.353	**0.683**													
	Factor 4	f4_y	-0.349	0.597	**0.666**												
Staff Version 1	Factor 2	f2_y	0.068	-0.085	-0.085	**0.744**											
	Factor 3	f3_y	0.026	-0.111	-0.084	0.162	**0.333**										
	Factor 4	f4_y	0.054	-0.078	-0.075	0.720	0.208	**0.697**									
Staff Version 2	Factor 2	f2_y	0.109	-0.085	-0.097	0.917	0.103	0.650	**0.710**								
	Factor 3	f3_y	0.018	-0.113	-0.086	0.532	0.870	0.465	0.337	**0.447**							
	Factor 4	f4_y	0.054	-0.078	-0.075	0.720	0.208	1.000	0.650	0.465	**0.697**						
Staff Version 3	Factor 2	f2_y	0.068	-0.085	-0.085	1.000	0.162	0.720	0.917	0.532	0.720	**0.744**					
	Factor 3	f3_y	0.039	-0.112	-0.086	0.113	0.950	0.156	0.071	0.804	0.156	0.113	**0.235**				
	Factor 4	f4_y	0.052	-0.075	-0.071	0.716	0.285	0.982	0.635	0.532	0.982	0.716	0.183	**0.695**			
Staff Version 4	Factor 2	f2_y	0.109	-0.085	-0.097	0.917	0.103	0.650	1.000	0.337	0.650	0.917	0.071	0.635	**0.710**		
	Factor 3	f3_y	0.023	-0.188	-0.091	0.535	0.829	0.450	0.336	0.972	0.450	0.535	0.836	0.478	0.336	**0.359**	
	Factor 4	f4_y	0.052	-0.075	-0.071	0.716	0.285	0.982	0.635	0.532	0.982	0.716	0.183	1.000	0.635	0.478	**0.695**

Table A. 12 Program Staff – Youth Correlation Matrix, with Cronbach's alphas, Guatemala Baseline

			Youth			Staff Version 1			Staff Version 2			Staff Version 3			Staff Version 4		
			Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4
			f2_y	f3_y	f4_y	f2_s1	f3_s1	f4_s1	f2_s2	f3_s2	f4_s2	f2_s3	f3_s3	f4_s3	f2_s4	f3_s4	f4_s4
Youth	Factor 2	f2_y	**0.708**														
	Factor 3	f3_y	-0.265	**0.698**													
	Factor 4	f4_y	-0.332	0.509	**0.706**												
Staff Version 1	Factor 2	f2_y	-0.017	0.026	0.033	**0.841**											
	Factor 3	f3_y	0.005	0.064	0.031	0.370	**0.363**										
	Factor 4	f4_y	-0.008	0.062	0.042	0.806	0.355	**0.732**									
Staff Version 2	Factor 2	f2_y	-0.016	0.004	0.008	0.953	0.279	0.736	**0.865**								
	Factor 3	f3_y	-0.003	0.072	0.054	0.640	0.911	0.584	0.479	**0.574**							
	Factor 4	f4_y	-0.008	0.062	0.042	0.806	0.355	1.000	0.736	0.584	**0.732**						
	Factor 2	f2_y	-0.017	0.026	0.033	1.000	0.370	0.806	0.953	0.640	0.806	**0.841**					

Staff Version 3	Factor 3	f3_y	0.011	0.076	0.037	0.357	0.946	0.309	0.286	0.851	0.306	0.357	**0.186**					
	Factor 4	f4_y	-0.011	0.060	0.040	0.788	0.448	0.983	0.704	0.655	0.983	0.788	0.350	**0.732**				
Staff Version 4	Factor 2	f2_y	-0.016	0.004	0.008	0.953	0.279	0.736	.000	0.479	0.736	0.953	0.286	0.704	**0.865**			
	Factor 3	f3_y	0.000	0.081	0.062	0.670	0.865	0.585	0.511	0.974	0.585	0.670	0.877	0.619	0.511	**0.504**		
	Factor 4	f4_y	-0.011	0.060	0.040	0.788	0.448	0.983	0.704	0.655	0.983	0.788	0.350	1.000	0.704	0.619	**0.732**	

Table A. 13 Program Staff – Youth Correlation Matrix, with Cronbach's alphas, Guatemala Endline

			Youth			Staff Version 1			Staff Version 2			Staff Version 3			Staff Version 4		
			Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4	Factor 2	Factor 3	Factor 4
			f2_y	f3_y	f4_y	f2_s1	f3_s1	f4_s1	f2_s2	f3_s2	f4_s2	f2_s3	f3_s3	f4_s3	f2_s4	f3_s4	f4_s4
Youth	Factor 2	f2_y	**0.756**														
	Factor 3	f3_y	-0.249	**0.701**													
	Factor 4	f4_y	-0.242	0.583	**0.720**												
Staff Version 1	Factor 2	f2_y	-0.009	-0.026	-0.002	**0.821**											
	Factor 3	f3_y	-0.078	0.087	0.046	0.387	**0.429**										
	Factor 4	f4_y	-0.012	-0.002	0.008	0.740	0.408	**0.684**									
Staff Version 2	Factor 2	f2_y	-0.014	-0.023	.001	0.950	0.332	0.667	**0.842**								
	Factor 3	f3_y	-0.057	0.054	0.030	0.648	0.913	0.598	0.502	**0.576**							
	Factor 4	f4_y	-0.012	-0.002	0.008	0.740	0.408	1.000	0.667	0.598	**0.684**						
Staff Version 3	Factor 2	f2_y	-0.009	-0.026	-0.002	1.000	0.837	0.740	0.950	0.648	0.740	**0.821**					
	Factor 3	f3_y	-0.014	0.094	0.056	0.356	0.942	0.379	0.300	0.854	0.379	0.356	**0.366**				
	Factor 4	f4_y	-0.004	0.002	0.006	0.734	0.459	0.977	0.657	0.667	0.977	0.734	0.402	**0.682**			
Staff Version 4	Factor 2	f2_y	-0.014	-0.023	0.001	0.950	0.322	0.667	1.000	0.502	0.667	0.950	0.300	0.657	**0.842**		
	Factor 3	f3_y	-0.073	0.055	0.036	0.658	0.863	0.600	0.508	0.970	0.600	0.658	0.886	0.619	0.508	**0.540**	
	Factor 4	f4_y	-0.004	0.002	0.006	0.734	0.499	0.977	0.657	0.667	0.977	0.734	0.402	1.000	0.657	0.619	**0.682**