Article:
Oulo, B., Thomas, M. and Sidle, A.A. (2026) Measuring agentic capacity: cross-cultural validation of the adolescent girls agency scale (AGAS). Humanities and Social Sciences Communications. ISSN: 2662-9992

https://doi.org/10.1057/s41599-026-07459-7

Humanities and Social Sciences Communications

Article in Press

https://doi.org/10.1057/s41599-026-07459-7

Measuring agentic capacity: cross-cultural validation of the adolescent girls agency scale (AGAS)

Received: 25 April 2025

Accepted: 22 April 2026

Cite this article as: Oulo, B., Thomas, M., Sidle, A.A. Measuring agentic capacity: cross-cultural validation of the adolescent girls agency scale (AGAS). *Humanit Soc Sci Commun* (2026). https://doi.org/10.1057/s41599-026-07459-7

Brenda Oulo, Matt Thomas & Aubryn Allyn Sidle

Measuring Agentic Capacity: Cross-Cultural Validation of the Adolescent Girls Agency Scale (AGAS).

Abstract

This paper presents the development and psychometric validation of the Adolescent Girls Agency scale (AGAS), a novel instrument specifically designed to measure agentic capacity among girls in East and Southern Africa. Addressing a critical gap in evaluation tools, the AGAS provides a structured framework for assessing girls' agentic capacity across four distinct domains: self-beliefs, gendered environmental beliefs, self-governance skills, and leadership skills.

The validation process employed a rigorous mixed-methods design, integrating qualitative cognitive interviews to ensure cultural relevance and clarity with item response theory analyses to assess psychometric properties, including reliability and validity, and to guide iterative scale refinement. Triangulation of statistical evidence on items requiring further review with stakeholder feedback ensured that the instrument reflected the lived experiences of girls in the target region while aligning with programmatic objectives.

The resulting final version of the AGAS provides a robust tool for pre–post evaluation of interventions aimed at strengthening girls' agentic capacity in East Africa. By offering domain-specific insights into how programs influence different dimensions of agency, the AGAS supports targeted program improvements and more informed decision-making. While three of the AGAS's four domains are hypothesized to be universal, further research is required to establish reliability and validity across diverse settings to support recommendations for adaptation in other comparable global contexts.

1. Introduction

Life skills programs play a vital role in adolescent development by nurturing a spectrum of psychosocial competencies and social support. However, evaluating the effectiveness of these programs presents significant challenges to researchers and practitioners alike (Duerden et al., 2012; Dupuy et al., 2018; NFER, 2023). The broad scope of life skills and the varied implementation approaches contribute to this difficulty. Furthermore, the lack of conceptual consensus and standardized instruments limits the comparability of program impacts, thereby hindering the ability to synthesize and identify best practices (Bernardo et al., 2023; Duerden et al., 2012). Existing tools often focus on isolated skills or program-specific outcomes, failing to capture the comprehensive impact of these programs. Moreover, many measures are designed within high-income country contexts, potentially limiting their applicability to the diverse structural and socio-cultural settings found in low- and middle-income countries (Berry et al., 2011).

The growing demand for culturally relevant and comprehensive outcome measures for life skills education and program evaluation has driven significant efforts in sub-Saharan Africa in recent years (Mugo et al., 2022; Ogunbiyi et al., 2025; Sidle et al, 2020). Guidelines for test or measure development frequently recommend a top-down approach, relying on expert panels to define concepts and formulate items, potentially overlooking bottom-up methods that could incorporate stakeholder insights on their lived experiences (AERA et al., 2014; Jagosh et al., 2012). Whereas, the Adolescent Girls Agency Scale (AGAS) was developed through a collaborative process, integrating stakeholder input from the conceptual framework development stage to item development, testing and consequently refinement (Sidle et al., 2020). Grounded in the Agentic Capacity Framework, which was co-conceptualized by researchers and community-based practitioners across Kenya, Uganda, Tanzania, and Rwanda, this

theory-driven scale is designed to assess girls' agentic capacity as a key outcome of life skills programs (Sidle & Oulo, 2023).

While the development of the AGAS has been detailed elsewhere (Sidle et al., 2020; Sidle & Oulo, 2023), this paper explores the cross-cultural validity and provides evidence of its item psychometric properties. The AGAS is currently used by a network of over 40 community-based, national, and international organizations to assess life skills programs across East, West and Southern Africa. Establishing both the validity of its use and the validity of its interpretation is a crucial foundation for the ongoing validation processes necessary to ensure the scale's accuracy and relevance as an assessment measure (S. G. Sireci, 2007). The use of reliable and valid assessment tools in program evaluation can enhance the design and implementation of more effective interventions, fostering evidence-based practices. Additionally, this paper highlights areas for future research, encouraging further adaptation of the AGAS for wider application in contexts with similar settings.

2. Theoretical Framework and Literature Review

Grounded in human capital and global education literature, this study examines the importance of a core set of competencies that are valuable across diverse jobs, industries, and life contexts (Catalano et al., 2019; Scales et al., 2015). Commonly referred to as "life skills," these competencies include self-awareness, self-management, decision-making, problem-solving, and interpersonal relationship skills. Increasingly recognized as critical determinants of successful youth transitions into the labor market (Heckman & Kautz, 2012; Heckman, James et al et al., 2006), these skills play a pivotal role in enabling young people to adapt, succeed, and thrive (Catalano et al., 2019).

Human development literature aligns with the definition of life skills as psychosocial and interpersonal competencies that empower individuals to make informed decisions, solve problems, communicate effectively, build relationships, and manage their lives

productively (Lerner et al., 2005; Ryan & Deci, 2000). While evaluations of life skills programs yield varied outcomes, they frequently demonstrate positive impacts on youth, particularly girls, in developing their agency, healthy relationships, and active citizenship (Temin and Heck 2022; Alvarado, G. et al., 2017; Catalano et al., 2019).

Albert Bandura's theory of the agentic self posits that individuals possess the capacity to influence their own functioning and the course of events through their actions (Bandura, 1989, 2006, 2023). Bandura emphasizes that agency is not solely an individual trait but is shaped through the dynamic interaction of personal, social, and environmental factors. His triadic reciprocal causation model illustrates this interplay, where behavior, personal attributes, social and environmental factors affect one another to shape human agency and outcomes (Bandura, 2023). At the core of this model is metacognition, encompassing intentionality, premeditation, self-reactivity, and self-reflection—key processes that allow individuals to regulate their thoughts, actions, and surroundings, thereby expressing their agency. Research further supports that agency can be enhanced through targeted, supportive interventions (Biglan et al., 2012; Durlak et al., 2011). This highlights the importance of designing policies and programs that are grounded in a deep understanding of local socio-cultural contexts to effectively promote agency among young people.

Drawing on Albert Bandura's theory of the agentic self, the Agentic Capacity Framework conceptualizes agentic capacity among girls as a multifaceted construct with four domains encompassing both skills and beliefs as shaped by contextual factors (Sidle & Oulo, 2023). This framework differentiates between skills and beliefs based on internal and external orientations. Internal beliefs encompass perceptions about the self, such as self-esteem and confidence, while external environmental beliefs reflect perceptions of one’s socio-cultural contexts, including gender norms and the extent to which they are malleable. Similarly, skills are categorized into internal self-governing skills, such as goal-setting and decision-making, and externally facing skills, such as communication,

public speaking and conflict resolution necessary for leadership and interpersonal relationships.

The standardization of multifaceted psychosocial outcome measures has been repeatedly advocated (Donald et al., 2017; Duerden et al., 2012; EASEL, n.d.). The Ecological Approaches to Social Emotional Learning (EASEL) Lab at Harvard offers an online resource for comparing organizational and region-specific frameworks to support this effort. (EASEL, n.d.). The Agentic Capacity Framework aligns with various life skills assessment frameworks and models in the EASEL database, including Lerner's Positive Youth Development framework (Lerner, 2009; Lerner et al., 2005) which highlights competence and confidence, among other key pillars.These frameworks share a common emphasis on shifting from the assessment of isolated skills to a broader evaluation of developmental assets and related domains. This focus on comprehensive youth developmental outcomes offers more practicality in crafting meaningful assessment tools designed to capture complex program outcomes.

However, the effective measurement of comprehensive developmental outcomes hinges on a clear and concise definition of those same outcomes. Agency, as an important example, has presented significant challenges to measurement due to its contextual variability and the lack of a unified definition and conceptual clarity (Hinson et al., 2021; Hitlin & Elder, 2006). Most frequently understood as the ability to act independently towards one's goals, agency encompasses both the potential to act and the actual exercise of that ability (Donald et al. 2020; Kabeer 1999). To refine this concept we differentiate between agentic capacity and agency or the ability to act versus action itself.

Although closely related, agency and empowerment are conceptually distinct. Agency, as already discussed, is an individual capacity, whereas empowerment denotes a transformative process that connects agency as a precondition of empowerment, with access to resources, and supportive contexts, in order to produce improved well-being (Kabeer, 1999; Drydyk, 2017). Empowerment therefore should entail removing

structural and social barriers that constrain girls' ability to exercise their agency (Chinman & Linney, 1998; Richardson, 2018; Zimmerman et al., 2019). Importantly, it must extend beyond the mere distribution of resources, which risks positioning girls as passive beneficiaries of development programs. Rather, empowerment should focus on creating pathways and spaces that enable girls to actively shape their own futures and support one another in driving collective transformation (Sen, 1999).

In resource-constrained environments, girls' agency is pivotal for challenging deeply ingrained gender norms that hinder women's economic independence (Hinson et al., 2021). Evaluation evidence supports the success of life skills programs in boosting girls' agency (Acharya et al., 2009; Dupuy et al., 2018; Marcus et al., 2017). For girls in challenging environments, agency is instrumental in transforming education into meaningful action that aligns with their personal aspirations and goals (Kwauk & Bragga, 2017; Lloyd & Hewett, 2009; Unterhalter, 2019). However, appropriately assessing this crucial construct has been a longstanding challenge to both researchers and practitioners.

2.1 Review of Existing Measures and Challenges with Measurement of Agency and Agentic capacity

Researchers approach agency from diverse perspectives. Some equate it with related constructs such as self-efficacy, autonomy, or self-determination, and subsequently assess it using similar measurement tools (Cavazzoni et al., 2022; Eteläpelto et al., 2013; Sutterlüty & Tisdall, 2019). Others, however, conceptualize agency as a distinct construct, albeit with varying definitions, such as "the socio-culturally mediated capacity to act" (Ahearn, 2001) or "the ability to define one's goals and act upon them"(Kabeer, 1999).

Research on human development and feminist economics further highlights the complexity of agency, emphasizing its interaction with the social, cultural and structural environment (Bandura, 2018; Kabeer, 2016; Russell, 1996). This complexity may

explain why many existing measures of girls' and women's agency utilize simple observable indicators of action such as the freedom of movement and household decision-making, while neglecting core agentic capacity (Cavazzoni et al., 2022; Ibrahim & Alkire, 2007; Yount et al., 2016). To overcome the challenges of contextual variability we argue that agency assessments should go beyond observational measures of agency-in-context to incorporate socio-culturally relevant measures of individuals' internal capacity to act or 'agentic capacity.' This approach would contribute towards the operationalization of the multi-dimensional scope of agency, provide a standardized framework for assessment, and help minimize current inconsistencies in measurement (Donald et al., 2020).

In pursuit of this goal, recent research has sought to clarify the internal attributes that define an individual's capacity to act. Sidle (2019) argues that positive self-belief and the skills needed to organize oneself into action are key components of agentic capacity (Sidle, 2019). This simplified yet broad understanding of agentic capacity provides a useful foundation for developing practical measures to assess agentic capacity as a desired outcome in youth development programs. The challenge of selecting relevant skills stems from the need to differentiate between universal skills and those specific to a particular context, as essential skills can vary depending on the social and structural environment. A systematic review of 34 studies elucidated this lack of standardization of skills assessed in measures of agency as every study adopted instruments evaluating varying skills except for multiple studies conducted by the same authors (Cavazzoni et al., 2022).

Greater consistency in agency measurements was noted in studies focused solely on women, which typically assessed expressions of agency such as household decision-making and mobility (Cavazzoni et al., 2022). When examining agency in youth, numerous studies utilized adapted measures originally designed to assess related constructs, including autonomy and self-efficacy (Berhane et al., 2019; Beyers et al., 2003; Chen et al., 2001), while others—particularly in high-income settings—developed

novel assessment tools that sometimes captured the agency of specific marginalized groups (Bentley-Edwards, 2016; Hitlin & Elder, 2006; Lautamo et al., 2021).

The scarcity of agency-specific measures for youth is highlighted by a systematic review which identified only a limited number of instruments suitable for adolescents (Gai et al., 2023). Additionally, instruments grounded in Bandura's triadic reciprocal causation model of agency only addressed distinct dimensions but failed to encompass all aspects of agency (Gai et al., 2023). This gap points to a theoretical limitation in existing tools, reflected in the narrow scope of their scales. Moreover, a significant gap remains in the development of agency-specific measurement tools for young people in low- and middle-income countries, where socio-cultural and economic contexts uniquely shape the development and expression of agency.

As already discussed, evaluating the impact of interventions designed to enhance agency is challenging without contextually relevant measurement tools that are theoretically grounded and demonstrate strong psychometric properties (Donald et al., 2020; Ibrahim & Alkire, 2007). The Agentic Capacity Framework, structured around four domains encompassing internal and external beliefs and skills, offered a culturally relevant foundation for addressing these measurement gaps through the development of the AGAS. Among these domains, self-belief, self-governance or management of oneself, and skills related to leadership or interpersonal relations are hypothesized to be universally applicable across diverse contexts (Dupuy et al.,2018). In contrast, environmental beliefs are hypothesized to be more context-dependent, varying even within the same cultural setting, as they are shaped by the surrounding opportunity structures.

This paper describes the cross-cultural adaptation of the AGAS and its validation as a pre-post program evaluation measure through a methodological study conducted across five culturally diverse yet economically similar countries namely Kenya, Tanzania, Rwanda, Uganda, and Malawi. Specifically, this study aims to achieve the following objectives:

Objective 1: To establish the content validity and psychometric properties of the AGAS items.

Objective 2: To utilize a mixed-methods and multi-stakeholder participatory process to refine the AGAS for contextual appropriateness across Kenya, Tanzania, Rwanda, Uganda, and Malawi.

Objective 3: Validate the AGAS for its intended use as a pre-post program outcome evaluation measure across the five study countries.

Description of the AGAS

The authors (2023) provide a comprehensive account of the development, validation, and interpretation of the AGAS, originally known as the AMPLIFY Agency scale. Initially consisting of 77 items, the scale was later refined to 60 items before being renamed the AGAS. It employs a five-point Likert scale, with response options ranging from 1 (strongly disagree) to 5 (strongly agree) for agreement-based items and from 1 (never) to 3 (always) for frequency-based items. These items are designed to evaluate four key domains of girls' agentic capacity:

- **Self-Beliefs:** 14 items related to personal beliefs about self-esteem and confidence in one's abilities e.g. I am satisfied with who I am as a person (SB1) and I'm a person of worth (SB2).
- **Environmental Beliefs:** 12 items that evaluate girls' perceptions of gender attitudes and norms, awareness of their rights, and their belief in their ability to influence their environment e.g. When the family cannot afford to educate all children, only boys should go to school (EB 6) and It is ok for a woman to earn more than her husband (EB 7).
- **Self-Governance Skills:** 18 items that evaluate skills related to coordinating one's behavior, thoughts, and abilities to take strategic action e.g. I am good at

setting goals for myself (SG 1) and When solving a problem in my life, I compare each possible solution with other solutions to find the best one (SG 4).

- **Leadership Skills:** 16 items that evaluate skills for effective interaction and communication, fostering positive relationships and collaborative environments e.g. I feel confident speaking in front of a big group (LS 12) and I am able to use what I have learned from my life to counsel others (LS 11).

Methods

Recognizing the importance of ongoing validity assessment (Arafat et al., 2016; S. Sireci & Benítez, 2023; S. G. Sireci, 2007), this methodological study employed a mixed-methods, cross-cultural validation approach, adapting the framework proposed by Arafat et al. (2016). The study integrated multiple methodological steps, including a content review process, qualitative cognitive interviews, quantitative scale data collection, and participatory tool refinement with stakeholders.

3.2 Content Review:

For assessments measuring abstract constructs related to individuals' beliefs and skills, content validity evidence is essential to ensure the instrument is used as intended (Delgado-Rico et al., 2012; S. Sireci & Faulkner-Bond, 2014). To establish the content validity of the AGAS, we assembled a panel of seven technical experts from diverse geographical contexts, including East Africa (Kenya and Tanzania), Malawi, India, and the United States. Their expertise in adolescent development, gender studies, life skills education, and psychometrics informed their assessment of each item’s clarity and relevance, both within its respective domain in the framework and in relation to agentic capacity as defined.

A mixed-methods approach was used to collect expert feedback. They provided quantitative ratings (1-4 scale), for both the clarity and relevance of each item where (1=Not clear/Not Relevant; 4=Very Clear/Very relevant) and provided qualitative notes

for any items rated 1 or 2. Inter-rater reliability was assessed using Gwet's AC1 coefficient. The obtained AC1 value was 0.72 (95% CI [0.60, 0.80]), indicating substantial agreement between raters (Gwet, 2014; Klein, 2018). Qualitative feedback on items with AC1 coefficients below .8 for clarity was carefully reviewed and items were revised based on synthesized suggestions. Items were dropped if coefficients for relevance did not meet a .7 or “substantial” threshold. Items were additionally dropped if the item was flagged by multiple reviewers as repeating other items on the instrument and if there were specific words or phrases identified by reviewers as difficult for adolescents to understand or challenging to translate into local languages. This process refined the original 60-item AGAS to a 48-item pool for further empirical evaluation, consisting of 10 self-belief items, 11 environmental belief items, 14 self-governance items, and 13 leadership items.

Translation and Back Translation

Translation is an initial step in cross-cultural adaptation, a process that requires developing instrument versions that are linguistically and culturally equivalent to the original, yet for different contexts. The AGAS was into multiple national languages across East Africa, including Kiswahili, Kinyarwanda, Chichewa, Luganda (and eight other Ugandan languages). Due to regional variations in Kiswahili between Kenya and Tanzania, separate Kiswahili translations were conducted for each country.

For each language, a professional translation firm was hired in each country (Malawi, Uganda, Tanzania and Rwanda) to conduct an initial translation of the instrument into the local language. Following cross-cultural validity guidelines, all translations were independently back-translated by a second professional translator, into English and subsequently reviewed by two members of the research team to ensure consistency with the original version (Arafat et al., 2016). Discrepancies in language or meaning identified by the research team upon review of the back-translation, were sent back to the translation firm for revision. Last, practitioners in each country whose organizations were part of the initial AGAS development process, convened to review the local

language version to ensure it conveyed the same cultural meaning as the original and that the language remained accessible and simple enough for an adolescent audience.

3.3 **Study Setting**

Data were collected in both rural and urban settings across Kenya, Tanzania, Rwanda, Uganda, and Malawi, reflecting the catchment areas of thirty-three collaborating Community Based Organization (CBOs). Nine of these organizations also participated in the co-creation of the Agentic Capacity Framework (Sidle & Oulo, 2023), while the remaining fourteen organizations were selected from a pool of over six hundred applicants across the five countries, to expand the geographical scope for data collection. The high number of applicant organizations highlights the growing demand for validated instruments to assess program outcomes effectively.

The final selection of organizations was purposive, and designed to ensure representation from diverse socio-cultural environments, including rural agricultural and fishing communities, remote pastoral and nomadic populations, as well as urban and peri-urban low- and middle-income areas. Table 1 lists the collaborating organizations in each country, while Figure 1 provides a visual representation of the geographical diversity covered. Each organization in Kenya, Tanzania and Rwanda collected approximately 200 observations each. In Malawi and Uganda there were fewer collaborating organizations and the amount per organization ranged from 200-800 based on geographic reach.

Table 1: Participating organizations and country representation

Country	Organizations
Kenya	1. TICAH 2. Akili Dada 3. Kakenya's Dream 4. Youth Changers Kenya 5. Msichana Empowerment Kuria 6. Northern Women Empowerment Initiative (NOWEI) 7. Dandelion Africa

	8. Chalbi Scholars Organization
Tanzania (Mainland & Zanzibar)	1. Tanzania Youth with New Hope In Life Organization (TAYONEHO) 2. Elle Peut Naidim (EPN) 3. Pamoja Youth Initiative 4. Secondary Education for Girls Advancement(SEGA) 5. The Girls Foundation of Tanzania 6. Jifundishe 7. Young Strong Mothers Foundation (YSMF) 8. Sawa Wanawake Tanzania (SAWA)
Uganda	1. Youth Fraternity for Change 2. Reach A Hand Uganda 3. I Profile Foundation 4. Girl Power Foundation Uganda 5. Elohim Development Association 6. Girl Up Initiative Uganda
Rwanda	1. Our Sisters' Opportunity 2. Impanuro Girls Initiative (IGI) 3. Glorious United for Rural Development 4. KOMERA 5. Streets Ahead Children's Centre Association (SACCA) 6. Learn Work Develop (LWD) 7. Rwanda Esther's Initiative (REI)
Malawi	1. CAMFED 2. Advancing Girls' Education in Africa (AGE Africa) 3. Maloto 4. Girls Empowerment Network

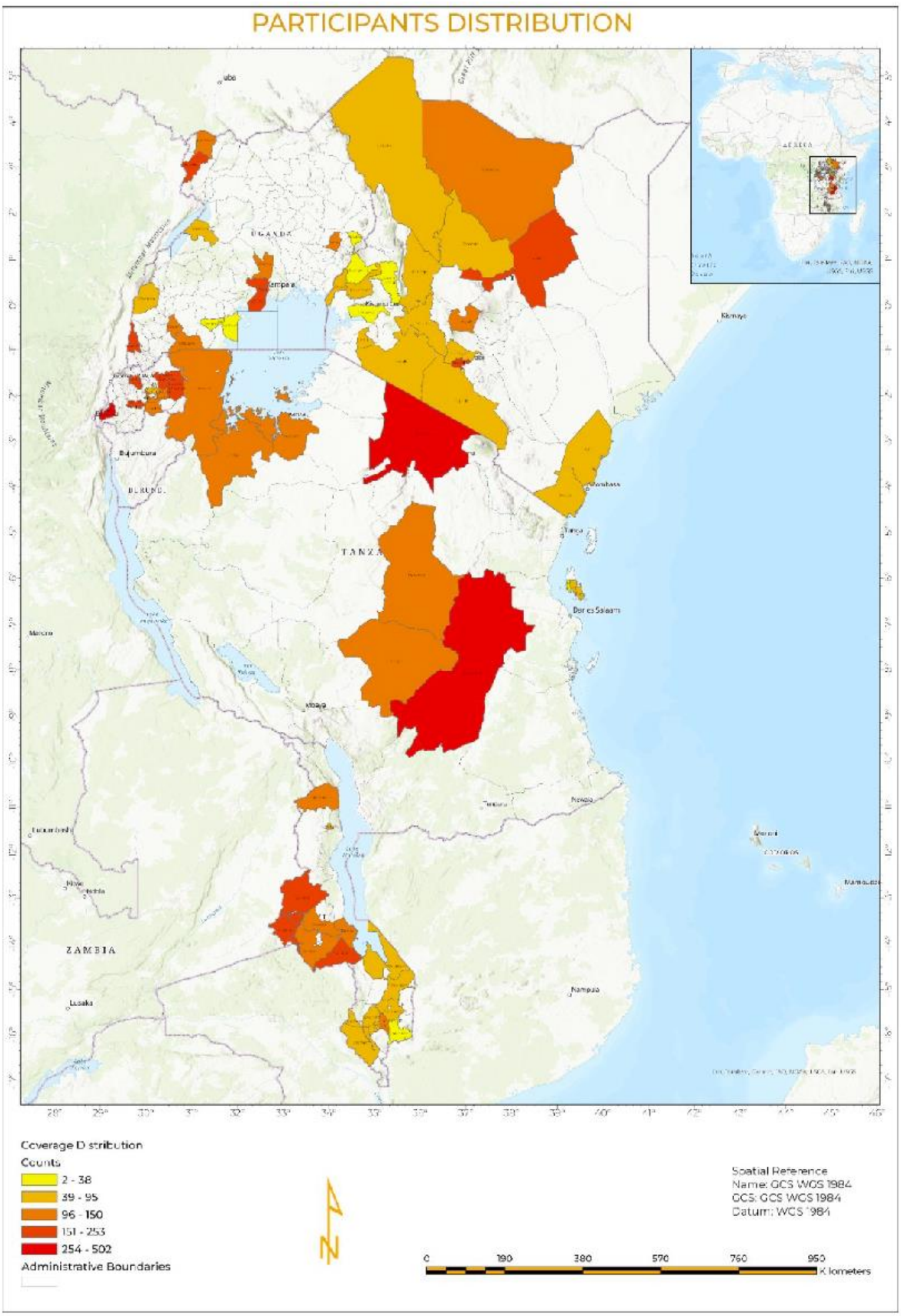
PARTICIPANTS DISTRIBUTION
UGANDA
BURUNDI
TANZANIA
ZAMBIA
Kampala
Bujumbura
Mombasa
Dar es Salaam
Kisaga
Nampula
Lusaka
Coverage Distribution
Counts
2 - 38
39 - 95
96 - 150
151 - 253
254 - 502
Administrative Boundaries
N
Spatial Reference
Name: GCS WGS 1984
GCS: GCS WGS 1984
Datum: WGS 1984
0
190
380
570
760
950
Kilometers

Figure 1: Regional areas highlighting organization catchment areas where the AGAS was administered.

3.4 **Participants**

This study aimed to reach 8,200 adolescent girls and young women (AGYW), with a stratified sample of 1,640 per country for quantitative data and 21 participants for cognitive interviews. Each CBO contributed between 200 and 800 participants, depending on the number of regions they represented. Participants were eligible if they were female, aged 10-24, provided consent, potentially enrolled in a collaborating organization's life skills program, and resided within the organization's catchment area. Individuals were excluded if they were male, outside the 10-24 age range, or declined to consent (participant or parental) in accordance with ethics approvals obtained for each country.

A pragmatic sampling frame was developed to capture geographical and cultural diversity in efforts to achieve meaningful country representation. It included 26 administrative regions: eight in Kenya, six in Tanzania, four in Uganda, five in Rwanda, and three in Malawi. Nine of the CBOs involved in co-creation of the Agentic Capacity Framework and an additional fifteen new CBOs were purposively selected based on their willingness to participate, geographic representation (within one of the administrative regions), expertise in girl-focused life skills programming, and ability to inform planned stakeholder consultations. This resulted in thirty-three participating CBOs across five countries, with higher representation in densely populated areas.

Next, purposive sampling was used to select between 200–800 participants per CBO, depending on the number of regions they represented. The samples included both enrolled program participants and non-enrolled individuals. To assess test-retest reliability, a subset of participants completed the scale twice within a two to four-week interval. Among scale takers, a further sample of 21 adolescent girls, including four with disabilities, participated in cognitive interviews across the five countries, with 3–6

participants per country. To ensure diverse perspectives, participants were grouped into three age categories: 10–13 years, 15–19 years, and 20–24 years.

Cognitive interviewing

In-depth interviews were conducted to understand how respondents interpreted and processed scale questions to help uncover problems with item wording, clarity, and ambiguity. Tourangeau's (1984) four-stage response model—comprehension, retrieval, decision-making, and response mapping—guided the cognitive interview data analysis process. Forthcoming reports will present detailed qualitative findings; however, it is important to highlight this process here to demonstrate the mixed-methods approach used to enhance the AGAS's cross-cultural adaptation and validity. Briefly, the findings indicated that the scale was generally well understood by most cognitive interview participants (73% of the time), with comprehension challenges most frequently observed among respondents aged 10–13 years and those with disabilities.

Data Analysis

The authors (2023) detailed the Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) via Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator in Mplus 8.4 (Byrne, 2012), that established the four-domain structure of the AGAS (formerly the AMPLIFY Agency Survey). Rather than replicating previous analyses, this study builds on them by confirming the four-domain structure through Item Response Theory (IRT)-based CFA, reliability testing while also evaluating item characteristics to strengthen validity evidence.

IRT Model

IRT is a methodology for analyzing responses such as those in the AGAS, in which individuals answer from a set of choices. Individuals are assumed to have trait levels (often referred to as ability) which determines the probability of responding in particular ways. In many cases, these individual level traits are associated with the probability of

answering a question correctly, though in applications such as the AGAS, the trait is associated with the probability of answering with any of the choices provided.

IRT enhances assessment precision by analyzing how individual items correspond to the underlying trait, allowing for more accurate and generalizable measurement. Unlike Classical Test Theory (CTT), which assumes equal item contributions to overall scores, IRT provides a more detailed analysis of item characteristics and their contributions to specific domains as well as participant abilities.

We used a multidimensional Graded Response Model (MGRM) to analyze the Likert-scale data obtained using the AGAS. The graded response model allows for the response variable to be considered as ordinal, meaning that we do not require a particular answer to be “correct” or that we assume that the distance between successive answers on the survey be equally spaced. Unlike the traditional unidimensional GRM, MGRM allows for the assessment of multiple, potentially correlated domains of the latent trait (Kehinde et al., 2022; Linden & Hambleton, 1997). Models were used to compare unidimensional (overall agentic capacity) and complex multidimensional models (specifying intercorrelation between domains), aligning with the theoretical and empirical support for a multidimensional agency construct. IRT analysis was conducted in R using the 'mirt' package (Chalmers, 2012) with the full-information maximum likelihood (FIML) estimator and plotting using the ggmirt package (Masur, 2022).

The model assumes ordinal response categories, with higher scores reflecting a greater level of the underlying construct. To ensure consistency, nine reverse-phrased AGAS items were recoded before analysis. IRT employs two key parameters to describe item functioning: discrimination and difficulty. Discrimination indicates how well an item differentiates between individuals with varying levels of a trait (e.g., self-beliefs). Items with lower discrimination capture a broader range of the trait, but with less precision, while those with higher discrimination focus on a narrower range of the trait, but with greater precision. Difficulty, or item location, reflects the level of the latent trait required

to transition from a specific response option to the next higher option (e.g. switching from “agree” to "strongly agree") (Linden & Hambleton, 1997). These traits can be combined into a set of characteristic curves for each item, indicating the probability of each response for a given trait value, and thus the most likely response for each trait value.

Before fitting the models, we analyzed the frequency distribution of item responses on the original 5-point scale and further reviewed cognitive interview data to understand how respondents utilized the scale. Findings revealed low response frequencies in certain categories. To address this, we symmetrically collapsed categories combining the two lowest response options ("strongly disagree" and "disagree"), the two highest ("strongly agree" and "agree"), while retaining the neutral category.

Model fit was assessed using a combination of overall fit indices (chi-square), comparative fit indices (CFI and TLI), and parsimony measures (RMSEA). Given the sensitivity of the chi-square statistic (Hu & Bentler, 1999; Chen, F. F, 2007), it was not relied upon exclusively. Instead, the following parameter guidelines were also used to determine model fit: RMSEA < 0.08 and CFI & TLI ≥ 0.90 (Brown, 2006; Hu & Bentler, 1999). Additionally, we confirmed that the Item Characteristic Curves (ICCs) for the collapsed categories remained smooth and continuous, as abrupt changes may indicate inappropriate collapsing (Linden & Hambleton, 1997) and continued to provide information for a range of scores while keeping sufficient discrimination.

Infit and Outfit Statistics

To identify whether specific items or respondents deviated from the expected response patterns of the instrument, infit and outfit statistics were computed for each domain. While related, these measures investigate different aspects of the fit.The infit statistic assesses whether responses from individuals whose ability (theta) is close to an item's difficulty are as expected, while the outfit statistic evaluates whether responses from individuals whose ability levels are far from an item's difficulty behave as expected. For both of these measures, values between 0.5 and 1.5 are considered acceptable (Wright & Linacre, 1994).

Measurement Invariance Procedures for Validating Cross-Country and pre-post Comparisons

Measurement invariance (MI) is a fundamental requirement for ensuring that scores from different groups can be meaningfully compared (Byrne, 2016; Vandenberg & Lance, 2000). MI confirms that the underlying concept being measured (the "latent construct") has the same meaning across those groups. Following the evaluation of individual item parameters and the removal of specific items (as detailed in the Results section), we analyzed the remaining data to provide preliminary evidence on whether scores from the shortened version could be meaningfully compared across countries and pre–post program participant groups. In this study, country was used as a proxy for cultural variability, with language versions aligned to respondents' countries. In Tanzania and Malawi, all participants completed the scale in Kiswahili and Chichewa, respectively, ensuring consistency between country and language. In Uganda, however, ethical requirements necessitated translation into more than six local languages, complicating language-based comparisons.

We followed the hierarchical series of increasingly restrictive steps as proposed by various authors (Byrne, 2016; Vandenberg & Lance, 2000) using the Lavaan R package implementing the diagonally weighted least squares estimator and listwise deletion for missing values. Beginning with configural invariance testing to assess whether the same basic factor structure holds across different country groups, we first established

baseline CFA models for each country separately, ensuring adequate model fit before fitting a simultaneous multigroup baseline model, allowing factor loadings to be freely estimated across groups without imposing any constraints. Subsequent models compare the fit of models holding additional parameters constant across groups.

The highest level of invariance, scalar invariance, requires constraining factor loadings and intercepts to be equal across groups providing the most robust foundation for comparing latent means across different groups. While full measurement invariance (all factors and intercepts constrained) is often difficult to achieve, partial invariance can still allow for meaningful comparisons between groups particularly when the number of non-invariant items is small (Byrne, 2016).

Assessment of Test-retest Reliability

To assess the temporal stability of the AGAS, we used IRT-based item analyses prior to scale shortening, followed by a longitudinal MI approach on the shortened scale (Pitts et al,1996). IRT strengthens test–retest reliability assessment by evaluating the stability of item parameters and respondent ability estimates (θ) across time points. Analyses were conducted using data from a subset of respondents who completed the scale twice within a two- to four-week interval, during which no intervention occurred. MI was tested across the full test–retest sample by jointly modeling the test and retest administrations to assess whether the scale's factor structure and model estimates remained stable over time, thereby ensuring that any observed score differences reflect true change in the underlying construct rather than measurement instability.

Results

AGAS data was obtained from a total sample of 8,208 participants included in the statistical analyses. Table 2 provides a breakdown by country and age group. Of these, 925 (11.27%) participated in the scale twice for test-retest reliability.

Table 2: Description of the AGAS participants

Number of participants (frequency)						
Age Category (Years)	**Kenya**	**Malawi**	**Rwanda**	**Tanzania**	**Uganda**	**Total (N)**
1. Below 14	413	424	182	445	587	2,051
2. 14 - 17	1,081	1,015	684	1,100	709	4,589
3. 18 - 20	165	243	539	65	274	1,286
4. Above 20	19	7	223	6	27	282
Total	**1,678**	**1,689**	**1,628**	**1,616**	**1,597**	**8,208**

Descriptive statistics for the AGAS items are detailed in Table 3, which reports the mean and standard deviation for each individual item to provide a granular overview of performance across the sample. Table 4 summarizes the internal consistency measures, presenting both McDonald's Omega and Cronbach's Alpha for each domain.

Table 3: Descriptive Properties of Domain-Specific AGAS Items

Item Number	Mean (SD)			
	SB	**SG**	**LS**	**EB**
1	4.52 (0.86)	2.5 (0.54)	2.35 (0.61)	3.94 (0.93)
2	4.28 (0.94)	2.37 (0.56)	2.52 (0.54)	4.08 (0.93)
3	3.53 (1.43)	2.44 (0.55)	2.56 (0.53)	4.14 (0.91)
4	4.47 (0.92)	2.35 (0.58)	2.35 (0.57)	4.20 (0.93)
5	3.72 (1.37)	2.33 (0.58)	4.49 (0.77)	4.09 (0.94)
6	4.32 (0.96)	2.38 (0.57)	4.24 (0.97)	4.73 (0.65)

7	3.93 (1.22)	2.40 (0.58)	3.86 (1.18)	4.62 (0.74)
8	3.79 (1.32)	4.13 (1.00)	3.86 (1.21)	2.36 (0.86)
9	4.22 (1.00)	2.66 (1.36)	4.08 (1.07)	2.56 (0.78)
10	3.91 (1.24)	3.85 (1.19)	4.31 (0.97)	2.79 (0.58)
11		4.00 (1.19)	4.51 (0.82)	2.83 (0.52)
12		4.34 (0.95)	4.09 (1.13)	
13		4.20 (1.12)	4.27 (0.98)	
14		3.97 (1.13)		

Table 4: Internal Consistency Reliability Estimates for AGAS Domains

Domain	McDonald's Omega	Cronbach's Alpha
SB	0.75	0.63
SG	0.82	0.74
LS	0.78	0.7
EB	0.77	0.61

Item Psychometric Properties

Item discrimination and difficulty varied across domains in the multidimensional GRM calibration, with slope estimates ranging from -0.07 to 3.49 (see table 4). Detailed psychometric properties for items in each domain are provided in the following sections, facilitating the identification of well-functioning items and those requiring further review. The final refined AGAS version was created based on these findings, combined with input from cognitive interviews and stakeholder consultations.

Self-Beliefs Domain

Infit and outfit statistics confirmed consistent response patterns and good item fit within the domain (Figure 2). With no items exhibiting unexpected behavior or contributing to measurement error, we proceeded to analyze individual item parameters and item characteristic curves (ICC). Item SB6 ("I can solve difficult problems if I try hard enough"), had the highest discrimination (slope = 1.71) in the self-belief domain, and contributed the maximum information within that domain (Table 5). In contrast, items SB3 (0.518) ("Sometimes I feel useless") and SB5 (0.589) ("I generally feel like a failure") showed low discrimination.

The corresponding ICCs for SB3 and SB5 (Figure 3) show that the green curves always remained low, indicating that the middle response option was never likely to be selected. Additionally, cognitive interviews revealed that the negative phrasing of these two items elicited feelings of sadness among respondents, suggesting their removal may be necessary. While item SB8 ("when faced with challenges I remain calm because I know I am adaptable") provided less overall information due to its lower location parameter, the likelihood of all response options being chosen suggests it may improve precision in measuring lower levels of self-belief compared to SB3 and SB5.

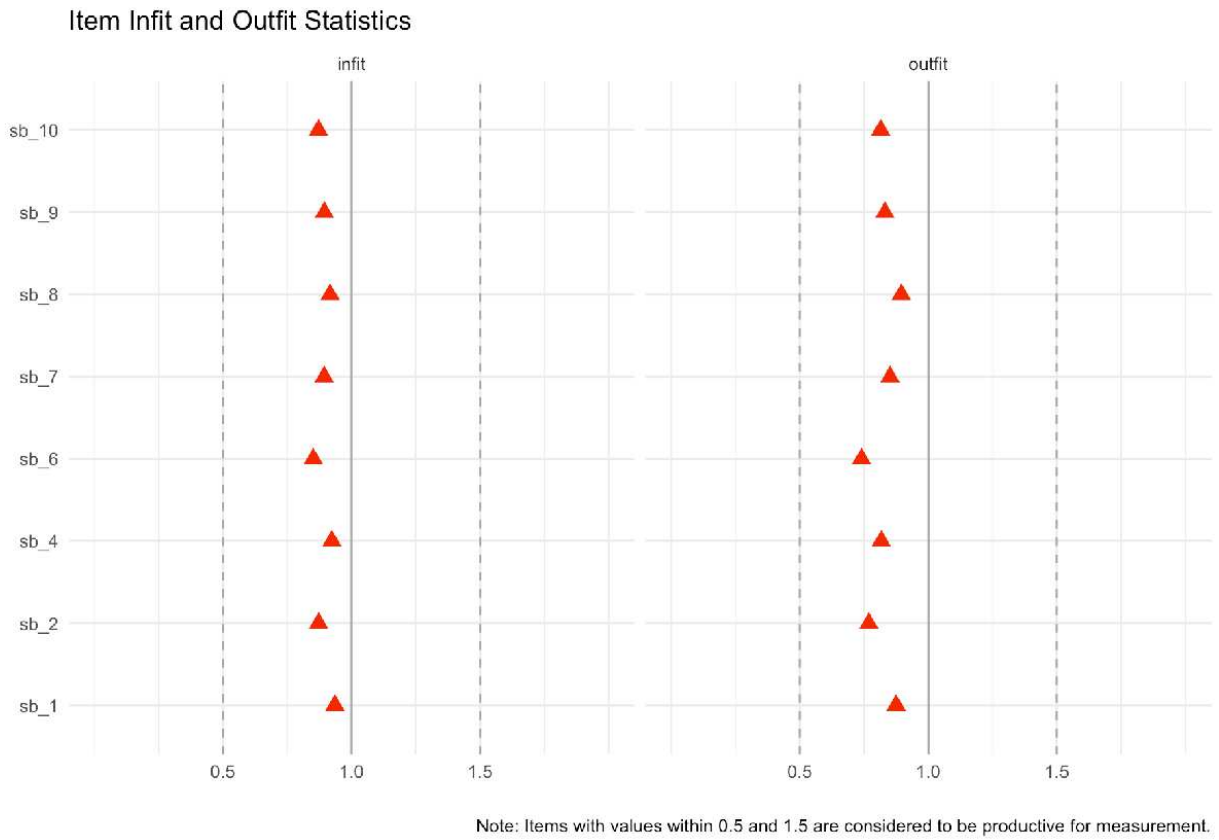

Figure 2: Infit and outfit for items in self-beliefs domain

Table 5: Item properties-self beliefs domain

Item	a	b1	b2	b3	b4
1. sb_1	1.4113	-3.3660	-2.7416	-2.1804	-0.8302
2. sb_2	1.5811	-3.1405	-2.2707	-1.5893	-0.1706
3. sb_3	0.5188	-4.0386	-1.9194	-1.0036	1.0839
4. sb_4	1.5735	-3.1229	-2.4370	-1.9002	-0.7362
5. sb_5	0.5894	-4.0107	-2.4392	-1.4180	0.5887
6. sb_6	1.7108	-2.8696	-2.1634	-1.6061	-0.2585
7. sb_7	1.0000	-2.7119	-1.8033	-1.1923	0.2238
8. sb_8	0.9186	-2.8797	-1.8152	-1.1642	0.4346

9. sb_9	1.3653	-3.2111	-2.3590	-1.6701	-0.0710
10. sb_10	1.3244	-2.5104	-1.6244	-1.0291	0.2141

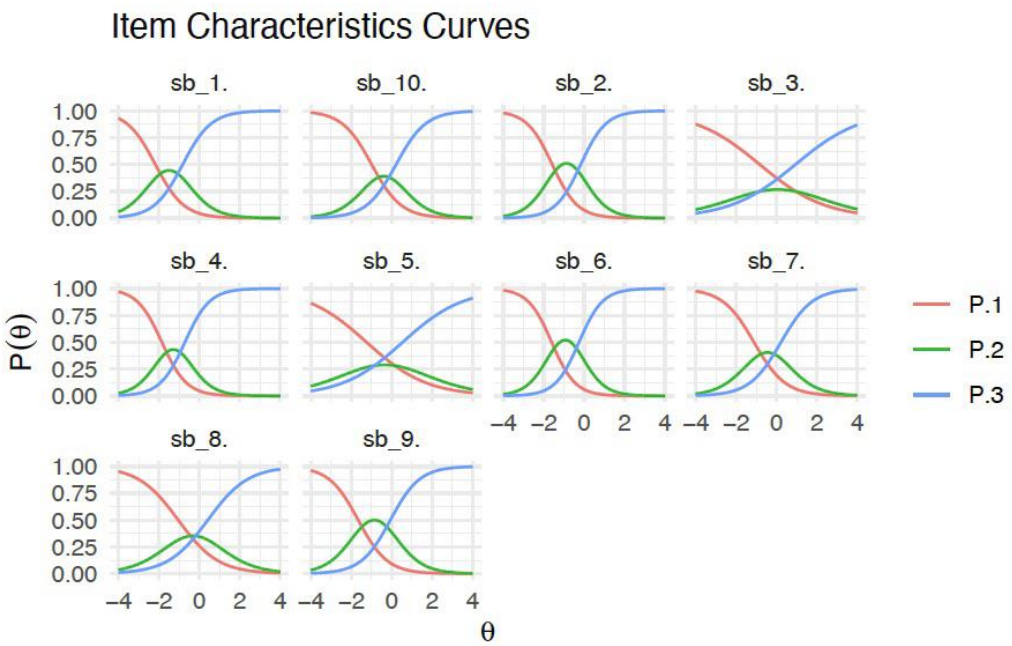

Figure 3: Item characteristic curves self beliefs domain

Self-Governance skills Domain

Response consistency and item fit were also assessed using infit and outfit statistics (Figure 4), which revealed no items behaving unexpectedly or contributing to measurement error relative to the other items in the domain. Within the self-governance domain, most items demonstrated low to moderate discrimination (0.86 to 1.32), indicating their ability to differentiate individuals with varying levels of the latent trait (Table 6). However, item SG9 (“it is sometimes hard for me to finish the tasks I start”) showed negligible discrimination, with a near-zero slope parameter suggesting limited utility in distinguishing individuals. Additionally, the item characteristic curves showed the scoring function were essentially linear (Figure 5), as a result, it was dropped from

the scale. Several items SG1 (“I am good at setting goals for myself”), SG2 (“when setting a goal I think about how much time it might take to achieve it”), and SG4 (“when solving a problem in my life I compare each possible solution with other solutions to find the best one”), along with SG3 (“when solving a problem I try to determine what caused the problem”) and SG12 (“I always think before I act”), exhibited similar discrimination levels. For parsimony, these items were flagged for further discussion to determine their retention based on their contribution to the construct and the overall integrity of the scale.

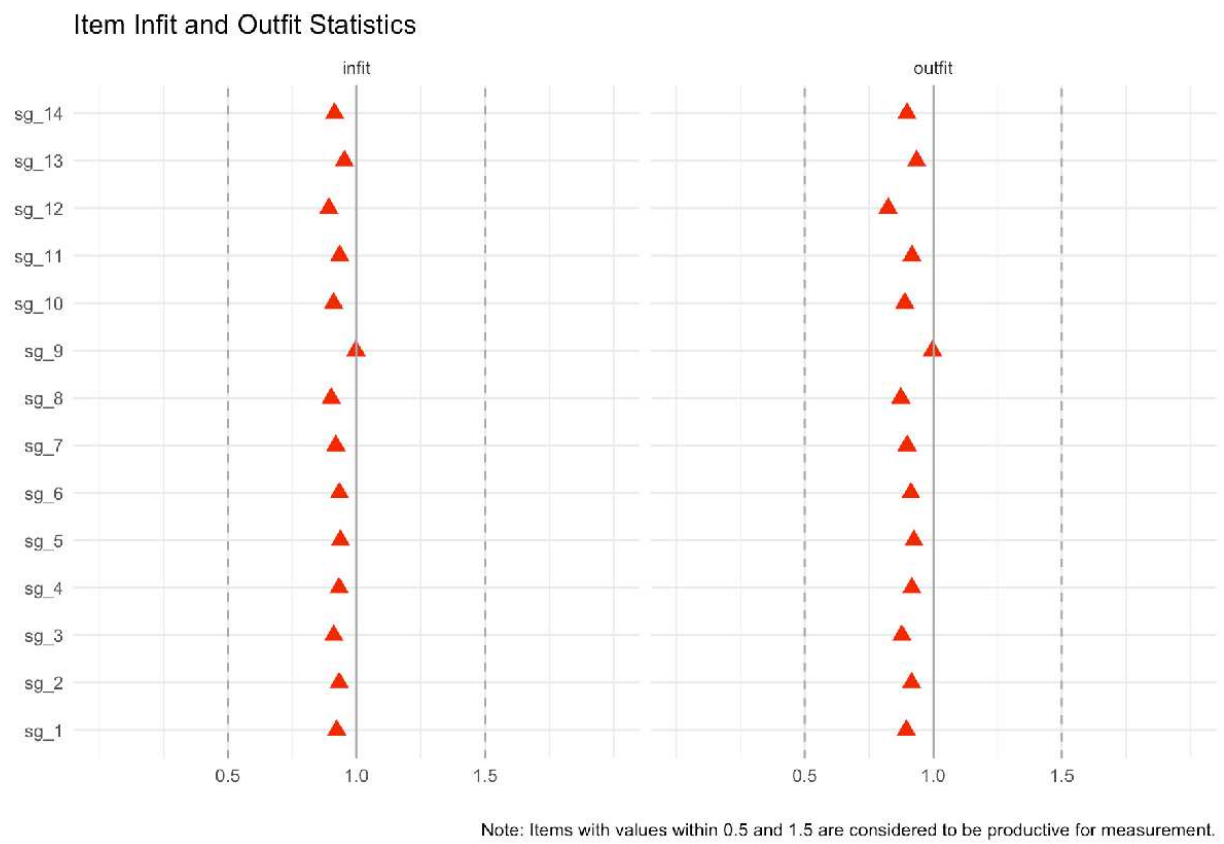

Figure 4: Infit & Outfit for items in the self-governance domain

Table 6: Item Parameters- self governance domain

Item	a	b1	b2	b3	b4
1. sg_1	1.1050	-4.0560	-2.8500	-1.1041	-0.1907
2. sg_2	1.1086	-3.5327	-2.2592	-0.8913	0.3359
3. sg_3	1.2710	-3.4362	-2.1218	-0.8511	0.0408
4. sg_4	1.1349	-3.0476	-1.8805	-0.6689	0.3639
5. sg_5	1.0570	-3.1975	-1.9506	-0.4783	0.4907
6. sg_6	1.0847	-3.3925	-2.0962	-0.8619	0.2697
7. sg_7	1.2607	-3.0466	-1.9150	-0.7537	0.1521
8. sg_8	1.1998	-3.5281	-2.5525	-1.6393	0.2183
9. sg_9	-0.0693	26.3988	11.1032	3.5634	-17.5944
10. sg_10	1.0614	-3.0809	-1.9876	-1.0840	0.5558
11. sg_11	0.9360	-3.4942	-2.1770	-1.4762	0.1572
12. sg_12	1.3224	-3.5611	-2.6650	-1.8858	-0.3452
13. sg_13	0.8644	-4.0105	-2.8119	-1.9900	-0.41959
14. sg_14	1.0211	-3.4306	-2.3028	-1.3820	0.4080

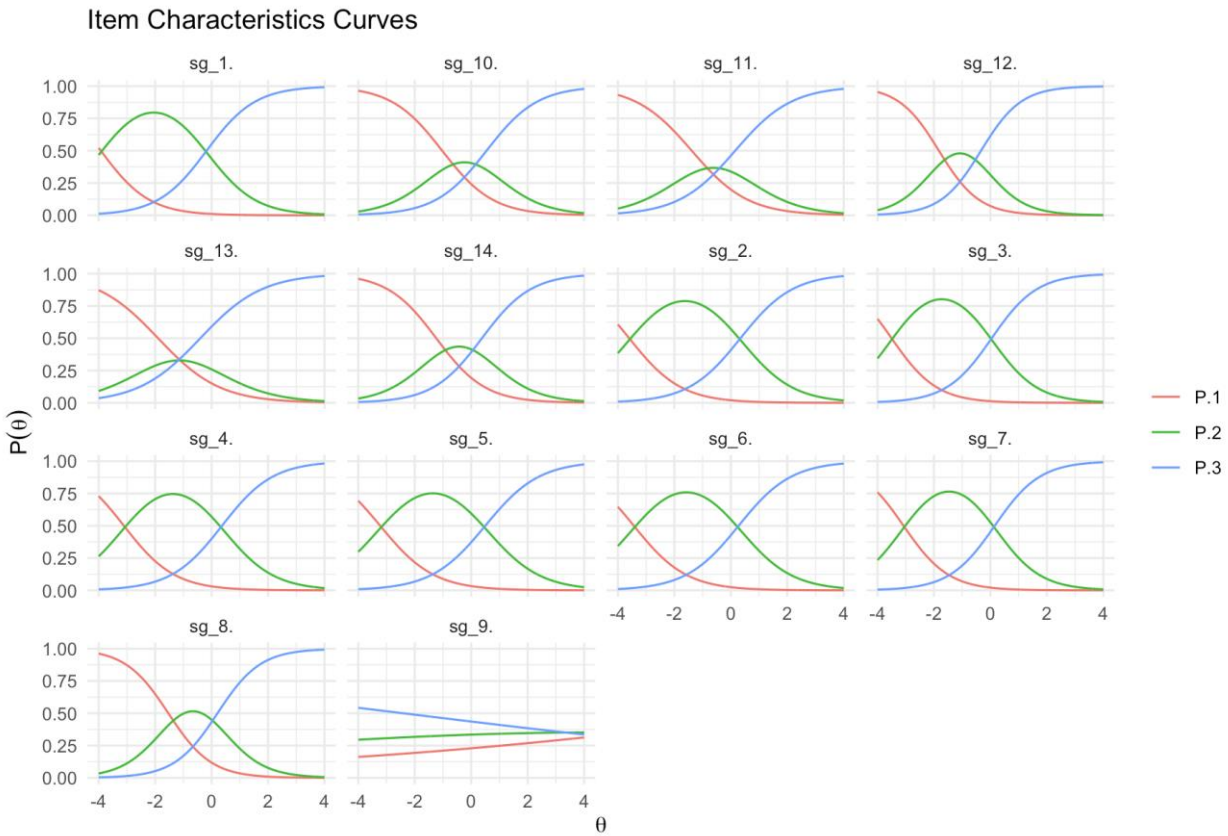

Figure 5: Item characteristic curves for items in self-governance domain

Leadership skills Domain

Consistent with the previous domains, infit and outfit statistics revealed no unexpected item behavior or measurement error within the leadership domain (Figure 6). However, item LS1 ("in a conversation I try to see the other person's point of view") exhibited the lowest discrimination (0.572) (Table 7) and was subsequently removed due to cognitive interview findings indicating participants misunderstood the phrase 'point of view' and had difficulty answering because of the varied nature of conversations. The question was also misinterpreted as the ability to convince someone. Additionally, LS2 ("I organize my thoughts before speaking"), LS3 ("I make sure I understand what another person is saying before I respond"), and LS4 ("When I need something, I am able to

express my needs to those around me") demonstrated lower discrimination and ICCs showing the lowest response option was rarely chosen (Figure 7), flagging them for further stakeholder discussion.

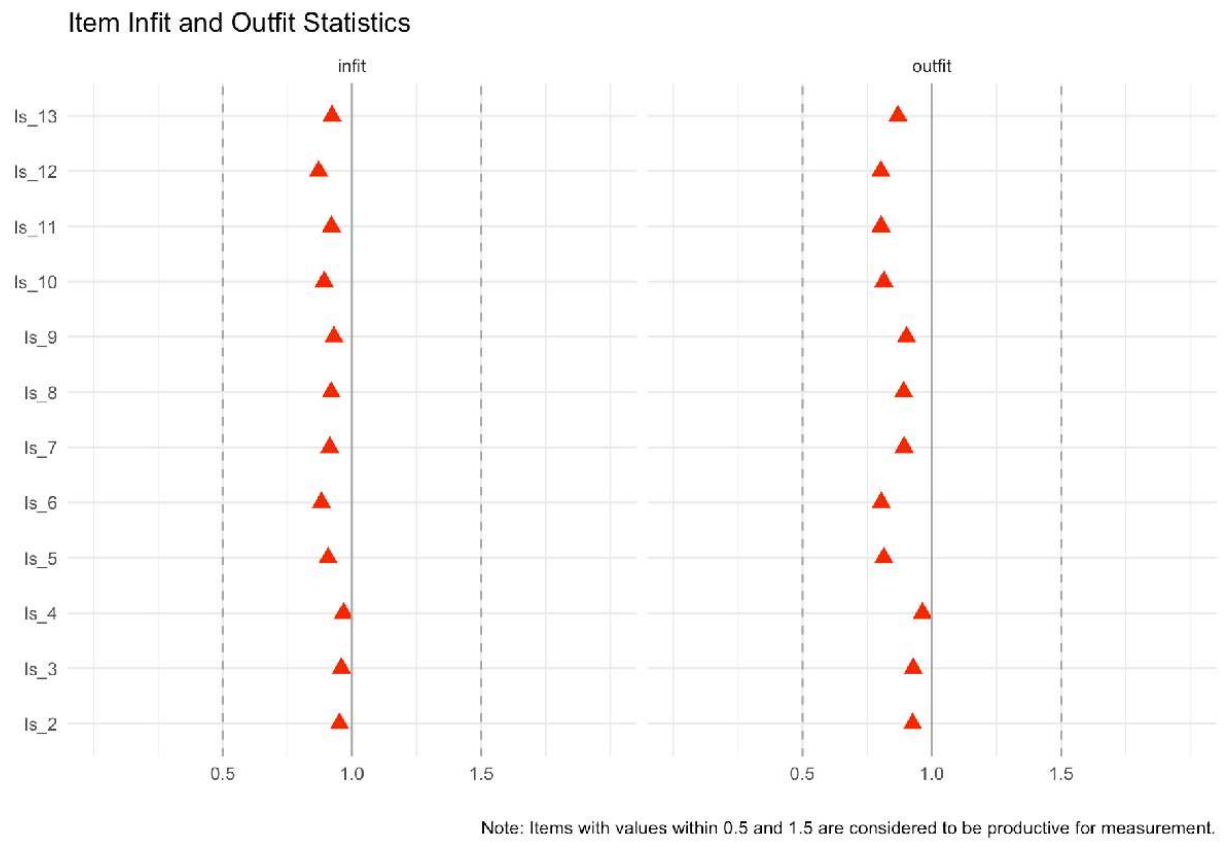

Figure 6: Infit & Outfit statistics for items in the leadership domain

Table 7: Item parameters-leadership domain

Item	a	b1	b2	b3	b4
1. ls_1	0.5725	-4.7823	-3.3745	-1.0971	0.5912
2. ls_2	0.9739	-4.4607	-2.9877	-1.4540	-0.2296
3. ls_3	0.9616	-4.7008	-3.1111	-1.7147	-0.4333
4. ls_4	0.7415	-4.3907	-2.5296	-0.5963	0.6075

5. ls_5	1.5124	-3.8707	-3.0570	-2.0868	-0.4743
6. ls_6	1.6136	-3.0759	-2.1975	-1.4953	-0.0535
7. ls_7	1.1491	-3.0272	-1.7572	-1.0076	0.5365
8. ls_8	1.0900	-3.0917	-1.7206	-1.0102	0.4669
9. ls_9	1.0793	-3.548	-2.3589	-1.5058	0.2777
10. ls_10	1.5322	-3.1207	-2.3249	-1.5669	-0.2865
11. ls_11	1.4301	-3.6201	-2.9308	-2.2284	-0.6440
12. ls_12	1.5305	-2.7360	-1.7839	-1.1951	0.0074
13. ls_13	1.2321	-3.5089	-2.6211	-1.8185	0.1398

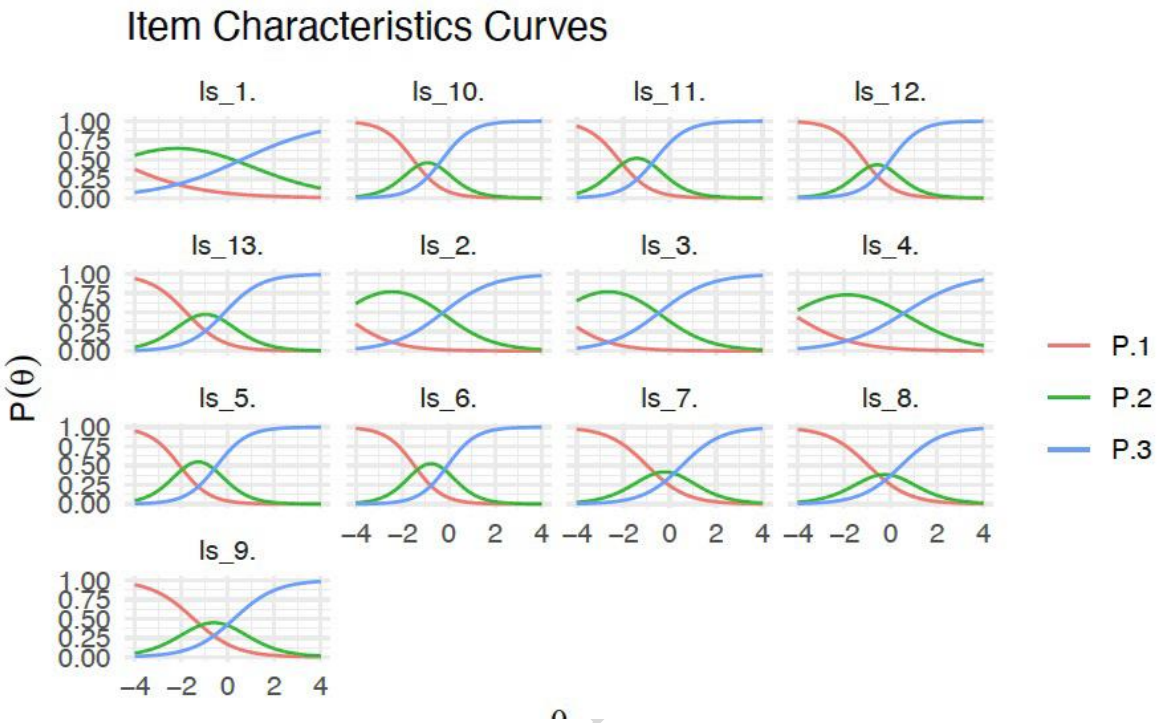

Figure 7: Item characteristic curves for items in the leadership domain

Environmental Beliefs domain

Within the environmental belief domain, items EB8 (0.3781) and EB9 (0.5739) exhibited low discrimination, while EB6 and EB7 had the highest slopes (Table 8), though potential misfit was noted in infit/outfit statistics (Figure 8). A closer review of items EB6 (“When the family cannot afford to educate all children, only boys should go to school”) and EB7 (“It is okay for a woman to have more education than her husband”) revealed that these items were perceived as ‘easy,’ with most respondents selecting the expected answers. The ICCs consistently showed a dichotomous response pattern, with the middle response option rarely chosen (Figure 9). To address potential limitations within this

domain, stakeholder discussions were planned to explore refinements to these items.

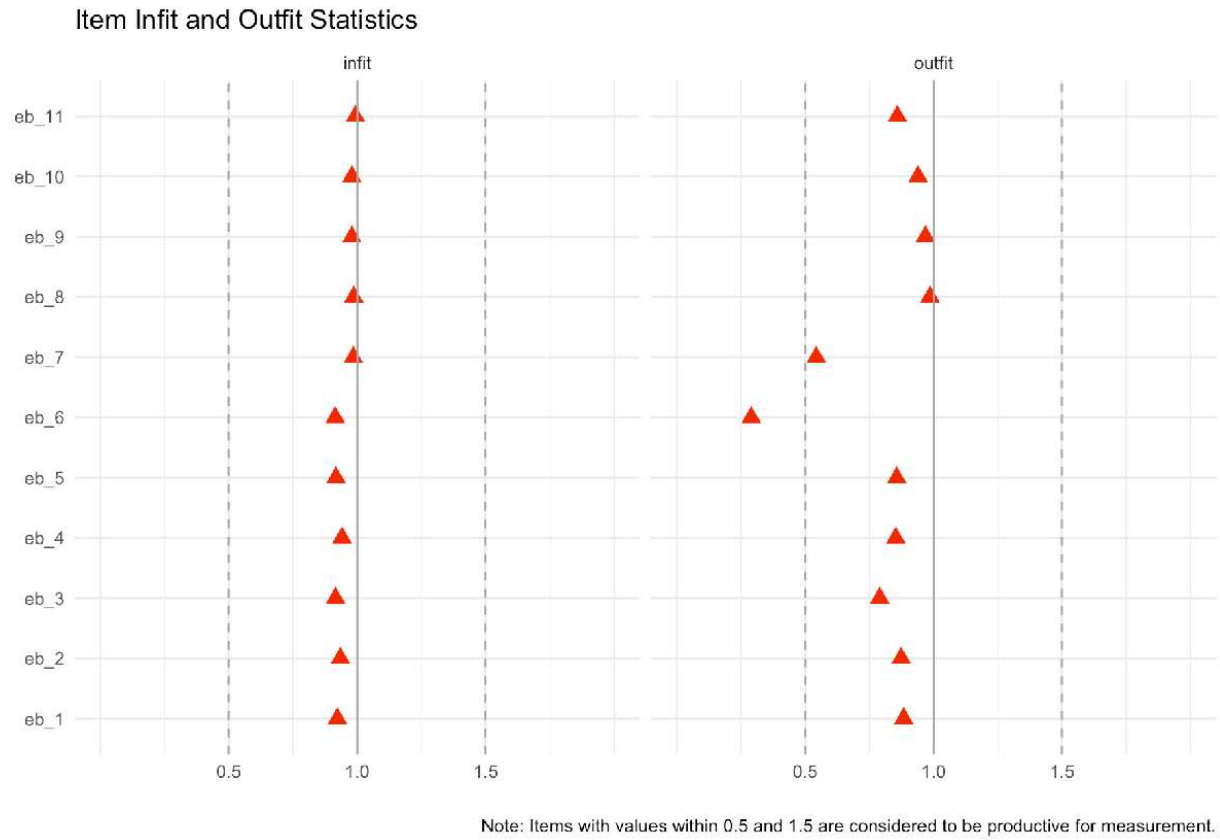

Figure 8: Infit and outfit statistics for items in the environmental beliefs domain

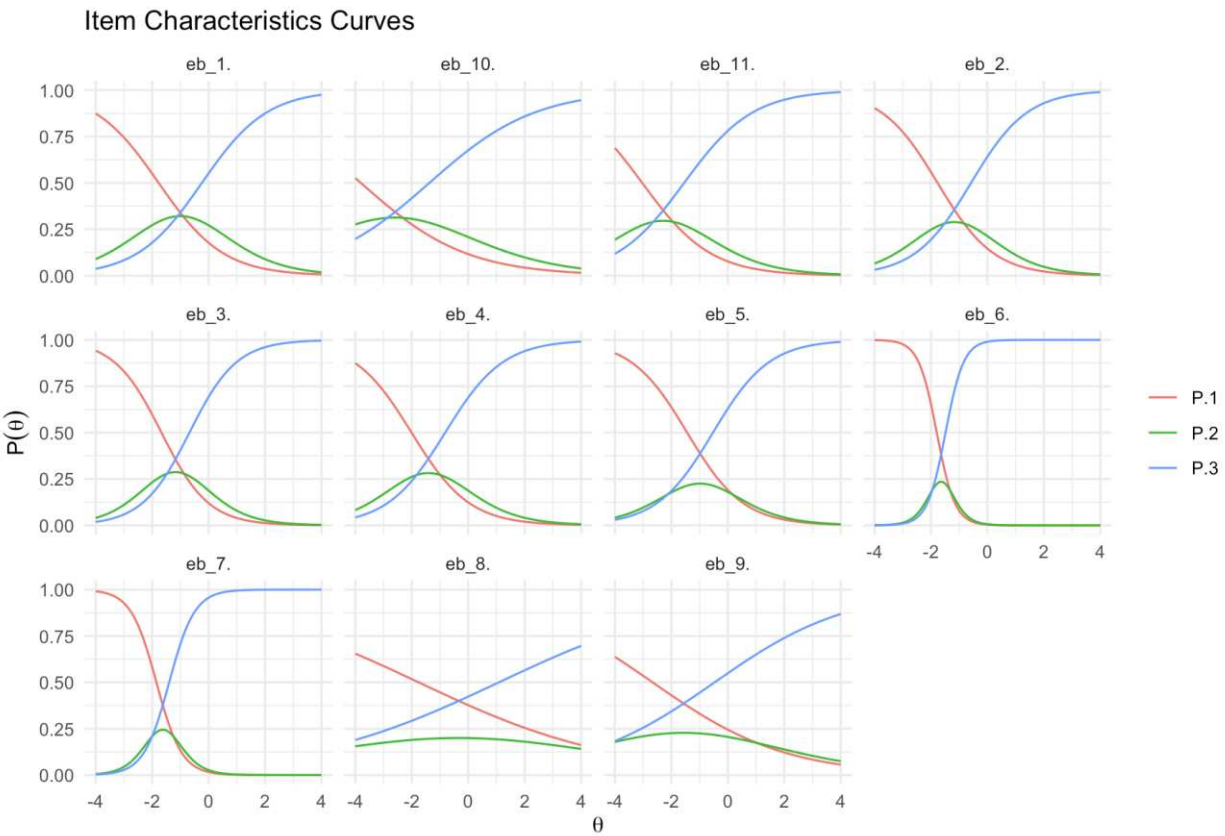

Figure 9: Item characteristic curves for items in the environmental beliefs domain

Table 8: Item parameters for Environmental Beliefs domain

Item	a	b1	b2	b3	b4
eb_1	0.8062	-1.8945	-1.8945	-0.3146	0.3896
eb_2	1.0232	-1.7610	-4.7610	-0.6307	0.0404
eb_3	1.1366	-1.7339	1.7339	0.7483	0.1120
eb_4	0.9615	-2.0397	2.0397	-0.8796	-0.3775
eb_5	1.0944	-1.3871	-4.3871	-0.5619	0.1198
eb_6	3.4896	-1.6977	1.6977	4.1559	1.1870
eb_7	2.4578	-1.7534	1.7534	-1.3464	-1.0298
eb_8	0.3781	-5.5579	5.5579	-3.0247	1.4174
eb_9	0.5739	-4.1782	4.1782	2.8680	2.0567

eb_10	0.8124	-4.3297	4.3297	2.3731	-2.6610
eb_11	1.1451	-3.5401	3.5401	2.8379	-2.3

In summary, across all four domains, location parameters (item discrimination) revealed a broad spectrum of underlying ability among respondents. High-discrimination items in each domain self-belief (SB2, SB4, SB6) self-governance (SG12, SG3) leadership (LS6, LS10, LS12) and environmental beliefs (EB6, EB7) are more effective at differentiating between individuals with different levels of the latent trait. When items are primarily endorsed by those with high trait levels ($\theta > 0$), it suggests they have high precision at the higher end of the trait distribution. For most AGAS items, the ICCs show varying probabilities of endorsing different response options at different levels of their respective domains of agentic capacity supporting the instrument's utility as a measure of evaluation.

Selection for a shorter refined AGAS Version

Based on findings that certain items did not function as expected and the need to enhance practicality by reducing respondent burden, the AGAS underwent a refinement process. This process integrated psychometric analyses, stakeholder input, and cognitive interview insights to ensure a robust revision. Items were retained based on their ability to provide maximum information about the underlying traits while maintaining the integrity of the construct.

Stakeholder input was gathered during a three-day onsite workshop in Nairobi in May 2024, where participants collaboratively reviewed findings from cognitive interviews and statistical analyses. Approximately 70 participants, including researchers, adolescent girls from all five countries and practitioners from each of the participating 33 organizations, and psychometricians, engaged in intensive discussions to discuss scale items that provided minimal information about relevant domain traits and consider their removal or retention.

Items that did not exhibit expected psychometric properties were discussed in plenary meetings and country-specific groups which further examined the equivalence of item translations against the original for linguistic and cultural appropriateness for their country contexts. Participants also engaged in a construct mapping exercise of the environmental beliefs and subsequently narrowed the items in this domain to be gender-specific. Based on these discussions, 15 items were dropped (Table 9), and three others were revised reflecting a triangulation of findings from cognitive interviewing, GRM analysis, and stakeholder input. The final 33-item AGAS instrument retained six self-belief, eleven self-governance, eight leadership, and eight environmental beliefs items, ensuring adequate representation of each domain's specific aspects as defined in the framework.

Table 9: Summary of items excluded from the final AGAS and the rationale for their removal or revision.

Domain	**Item(s)**	**IRT-Related Reasons for Item Removal**	**Qualitative Rationale for Item Removal**
Self-Beliefs (SB)	2,3,5,6	Low discrimination	Items 3 and 5 phrasing evoked negative emotions, stakeholder suggestion to omit
Self-Governance Skills (SG)	6, 9, 11	Low discrimination	Items 6 and 11 demonstrated universal poor comprehension across all cognitive interviews.
Leadership Skills (LS)	1, 2, 5, 7, 9	Low discrimination	Items 2 and 5 were flagged by stakeholders for having ableist connotations. Items 7 and 9 were removed based on girls' feedback regarding lack of relevance.
Environmental Beliefs (EB)	1, 2, 8	Items 6 and 7 were retained following revision, despite initial poor infit/outfit statistics.	Scope of the items was too general, failing to capture the domain's intent regarding gendered constraints. Item 9 was revised due to stakeholder

			feedback

Measurement Invariance findings

The combined model exhibited acceptable fit indices (Table 10), supporting configural invariance. This finding indicates that the shortened AGAS maintains the same underlying conceptual structure across all groups, allowing us to proceed with subsequent invariance testing. Table 11 represents fit indices for the different countries for metric invariance. When compared to the baseline model, revealing that metric invariance was not achieved ($X^2(264) = 1064.1$, $p < 0.01$), precluding direct cross-country comparisons of latent means and correlations, as discussed previously.

Table 10: Model Fit Indices for full sample

Model	CFI	RMSEA	SRMR
Configural	0.952	0.046	0.073
Metric	0.943	0.047	0.075
Scalar	0.924	0.053	0.076

Table 11: Metric Invariance Model Fit Indices for Different Country Groups

Country	**TLI**	**CFI**	**RMSEA**
All countries combined (configural)	.949	.953	.043
1.Kenya	.943	.948	.043
2.Tanzania	.947	.951	.047
3.Uganda	.951	.956	.046
4.Malawi	.950	.954	.049
5.Rwanda	.947	.951	.044

Test-Retest Reliability findings

IRT analysis showed that item parameters were stable across the two administrations of the survey. The following potentially problematic items were identified at the domain level: SB domain (items 3, 5, and 8), SG domain (item 9), and LS domain (items 1 and 9). The EB domain showed no problematic items. This suggested that most of the items on the AGAS were already reasonable for test-retest purposes, and the potentially problematic items were noted for discussion with stakeholders. MI model fit indices for test-retest groups are provided in Table 12. Both configural and metric invariance was achieved as there was no significant different in fit between the configural and metric models ($X^2(66) = 64.63$, $p = 0.525$), as was scalar invariance ($X^2(29) = 40.141$, $p = 0.082$).

Table 12: Fit indices of test-retest reliability models

Model	CFI	RMSEA	SRMR
Configural	0.954	0.043	0.063
Metric	0.954	0.041	0.063
Scalar	0.954	0.041	0.063

Preliminary Evidence of Validity for Using the AGAS in Pre–Post Evaluation

TLI, CFI, and RMSEA values of 0.950, 0.954, and 0.043 respectively from a multi-group model suggest that the factor structure holds up well for both pre and post measurements. Configural invariance was reasonable to assert when assessing model fit of the pre and post groups against the full model. Testing the impact of constraining factor loadings reveals that metric invariance is also established (X2(66) = 64.631, p = 0.525). Furthermore, testing the restriction of equal intercepts also indicates that scalar invariance is achieved (X2(29) = 40.141, p = 0.082). These preliminary findings indicate

that comparisons of pre-post mean scores obtained following an intervention are appropriate. While the scale has been refined, additional research with pre-post groups is required to confirm its validity in measuring intervention-related changes in agentic capacity.

Discussion

To enhance the precision, validity of use, and cross-cultural applicability of the AGAS as a measure of girls' agentic capacity in program evaluation across five sub-Saharan African countries, this methodological study employed a robust mixed-methods approach. We integrated statistical evidence derived from IRT, a powerful tool for scale development and validation, with insights into respondents' cognitive processes, aligning with established frameworks for cultural adaptation and psychometric validation (Arafat et al., 2016). Furthermore, our study is strengthened by the use of a participatory approach to scale refinement, actively incorporating the perspectives and expertise of local stakeholders and girls. This collaborative approach ensured that the AGAS was carefully refined to enhance its relevance and applicability across diverse contexts, thereby strengthening its validity of use.

Existing measures of women's and girls' agency frequently focus on observable indicators such as household decision-making and mobility, neglecting the critical internal and psychological dimensions (Cavazzoni et al., 2022; Donald et al., 2020). This study addresses this gap by offering enhanced conceptual clarity and precise measurement applying a multidimensional framework of these internal aspects, termed 'agentic capacity'. The Agentic Capacity Framework encompasses beliefs about self and environment, as well as skills related to self-governance and interpersonal leadership, and aligns with broader research advocating for conceptual clarity of models assessing individual agency. Building on a World Bank review (Donald et al, 2020) that highlights 'sense of agency' as a key dimension defined by perceived control and ability, this study enhances clarity around these internal aspects of agency and provides

empirical support for the AGAS as a valid and reliable measure for evaluating programs where strengthening agentic capacity is a central goal.

Psychometric analysis of the AGAS items revealed their ability to effectively discriminate between low and high levels of agentic capacity domains. For complex MGRM models, accurate item parameter estimation is enhanced by longer tests (e.g., 40 items) and large sample sizes (Kehinde et al., 2022) both of which were features of our model. Comparing tests administered within a two- to four-week interval demonstrated the stability of individual item parameters and person ability estimates, indicating that the AGAS measures a construct that remains stable without intervention or change minimizing measurement error (AERA et al., 2014).

Evidence of partial measurement invariance supports the cross-cultural adaptation of the AGAS, with configural invariance achieved but not scalar or metric invariance across countries (Arafat et al., 2016; Byrne, 2016). While configural invariance validates the fundamental four-domain structure of agentic capacity across all contexts, the subsequent failure to achieve scalar invariance limits direct comparisons of AGAS mean scores between countries. This outcome is unsurprising, given our hypothesis that three domains are universal while the environmental beliefs domain is context-specific. Given that non-invariance is a common finding in cross-cultural studies of complex constructs (Davidov et al., 2018; Leitgöb et al., 2022) the established configural invariance remains strong evidence supporting the internal validity of the scale and permitting meaningful within-country comparisons. Future research on the AGAS should shift the focus of non-invariance testing from a purely methodological hurdle to a sociological inquiry. This requires moving beyond aggregate scale comparisons to individually investigate the domains. Specifically, comparison of the hypothesized universal domains against the context-specific domain (environmental beliefs) is needed to substantively understand the sources of differences in scale measurements across countries.

Designed for contextual adaptability, the AGAS incorporates three domains (Self-Beliefs, Self-Governance, and Leadership) that reflect broadly universal skills for

adolescents, providing a stable, cross-contextual foundation. Complementary evidence from cognitive interviews and stakeholder consultations supported the hypothesis that these three domains demonstrate wide adaptability. In contrast, the environmental beliefs domain, is more gender-specific capturing a girl's perceptions of the gendered norms in her environment and her belief in shaping them through action, was identified as highly context-specific, varying significantly across community, regional, and national levels.

This finding is congruent with existing research highlighting the profound influence of socio-cultural and structural environments on agency (Ahearn, 2001; Bandura, 2018; Kabeer, 1999). Accordingly, the AGAS offers a gender-responsive scale which can be adapted for other contexts, future adaptation efforts for the AGAS should primarily focus on refining the environmental beliefs domain to accurately capture girls'perceptions regarding the rigidity of their local gendered norms. Our study highlights the utility of IRT in combination with participatory processes which allowed us to refine the AGAS into a shorter version by integrating multiple forms of validity evidence while engaging stakeholders to ensure the integrity of the defined constructs. Further research is needed to assess the refined AGAS's suitability for cross-country comparisons, either in total scores or restricted to specific subscores, given the extensive multi-country stakeholder input in its development. The psychometric findings are limited by incomplete evidence of construct validity across all language versions. Subsequent research should prioritize rigorous language validation to ensure conceptual equivalence. We recommend further investigating the measurement invariance of the hypothesized universal domains within the Agentic Capacity Framework and the shortened AGAS across diverse country contexts.

In line with validation guidelines, additional validity evidence for the AGAS can be generated by examining convergent validity through comparisons with established instruments that measure similar constructs (S. Sireci & Benítez, 2023). The shortened AGAS represents a rigorously developed instrument supported by encouraging

preliminary evidence of cross-cultural validity. It is specifically designed to assess girls' agentic capacity within the unique socio-cultural contexts of East and Southern Africa, providing a foundation for further validation across diverse populations. Our review found that most comparable measures were organization-specific, designed to evaluate specific life skills programs (Gandara, 2024) or predominantly measure the manifestations of agency (Ogunbiyi et al., 2025). The refined AGAS fills this critical gap to evaluate the underlying agentic capacity as there is growing recognition of agency as a key driver and mediator of related health and education outcomes (Raj et al., 2024).

This paper supplements previously reported validity evidence (authors, 2023), positioning the AGAS as a possible benchmark for future convergent validity studies with emerging measures of agentic capacity. Ongoing studies are further examining these relationships and the results are forthcoming (Gandara et al., 2024).

Data Available on Request

Data supporting this study's findings are available to reviewers upon request from the corresponding author. However, due to third-party restrictions, the data are not publicly accessible.

References

Acharya, R., Kalyanwala, S., Jejeebhoy, S., & Nathani, V. (2009). *Broadening girls' horizons: Effects of life skills education programme in rural Uttar Pradesh*. Population Council. https://doi.org/10.31899/pgy15.1002

AERA, APA, & National Council on Measurement in Education (Eds.). (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association. https://www.apa.org/science/programs/testing/standards

Ahearn, L. M. (2001). Language and Agency. *Annual Review of Anthropology*, *30*, 109–137.

Alvarado, G., Skinner, M, Plaut, D, Moss, C., Kapungu, C, & Reavley, N. (2017). *A systematic review of positive youth development programs in low- and middle-income countries*. USAID. https://pdf.usaid.gov/pdf_docs/PA00MR58.pdf

Arafat, S., Chowdhury, H., Qusar, M., & Hafez, M. (2016). Cross Cultural Adaptation and Psychometric Validation of Research Instruments: A Methodological Review. *Journal of Behavioral Health*, *5*(3), 129. https://doi.org/10.5455/jbh.20160615121755

Bandiera, O. B., Niklas Burgess, Robin Goldstein, Markus Gulesci, Selim Rasul, Imran Sulaiman, Munshi. (2017). *Women's Empowerment in Action*. World Bank. https://doi.org/10.1596/28282

Bandura, A. (1989). Human agency in social cognitive theory. *The American Psychologist*, *44*(9), 1175–1184. https://doi.org/10.1037/0003-066x.44.9.1175

Bandura, A. (2006). Towards a Psychology of Human Agency. *Perspectives on Psychological Science*, *1*(2), 164–180.

Bandura, A. (2023). Social Cognitive Theory: An Agentic Perspective on Human Nature. John Wiley & Sons, Inc.

Bentley-Edwards, K. L. (2016). Hope, Agency, or Disconnect: Scale Construction for Measures of Black Racial Cohesion and Dissonance. *Journal of Black Psychology*, *42*(1), 73–99. https://doi.org/10.1177/0095798414557670

Berhane, Y., Worku, A., Tewahido, D., Fasil, N., Gulema, H., Tadesse, A. W., & Abdelmenan, S. (2019). Adolescent Girls' Agency Significantly Correlates With Favorable Social Norms in Ethiopia—Implications for Improving Sexual and Reproductive Health of Young Adolescents. *Journal of Adolescent Health*, *64*(4, Supplement), S52–S59. https://doi.org/10.1016/j.jadohealth.2018.12.018

Bernardo, A. B. I., Albert, J. R. G., Vizmanos, J. F. V., & Muñoz, M. S. (2023). *Toward Measuring Soft Skills for Youth Development: A Scoping Study*. https://www.econstor.eu/bitstream/10419/284627/1/pidsdps2328.pdf

Berry, J. W., Ype, H. P., Seger M, B., Athanasios, C., & David L, S. (2011). *Cross-Cultural Psychology | Third Edition*. Cambridge University Press. https://www.cambridge.org/highereducation/books/cross-cultural-psychology/6CB886EC3F28EED1269411F111723749#overview

Beyers, W., Goossens, L., Vansant, I., & Moors, E. (2003). A Structural Model of Autonomy in Middle and Late Adolescence: Connectedness, Separation, Detachment, and Agency. *Journal of Youth and Adolescence*, *32*(5), 351–365. https://doi.org/10.1023/A:1024922031510

Biglan, A., Flay, B. R., Embry, D. D., & Sandler, I. N. (2012). The Critical Role of Nurturing Environments for Promoting Human Wellbeing. *The American Psychologist*, *67*(4), 257–271. https://doi.org/10.1037/a0026796

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. The Guilford Press.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming* (pp. xvii, 412). Routledge/Taylor & Francis Group.

Byrne, B. M. (2016). Multigroup comparisons: Testing for measurement, structural, and latent mean equivalence. In *The ITC international handbook of testing and assessment* (pp. 377–394). Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0026

Catalano, R. F., Skinner, M. L., Alvarado, G., Kapungu, C., Reavley, N., Patton, G. C., Jessee, C., Plaut, D., Moss, C., Bennett, K., Sawyer, S. M., Sebany, M., Sexton, M., Olenik, C., & Petroni, S. (2019). Positive Youth Development Programs in Low- and Middle-Income

Countries: A Conceptual Framework and Systematic Review of Efficacy. *Journal of Adolescent Health*, *65*(1), 15–31. https://doi.org/10.1016/j.jadohealth.2019.01.024

Cavazzoni, F., Fiorini, A., & Veronese, G. (2022). How Do We Assess How Agentic We Are? A Literature Review of Existing Instruments to Evaluate and Measure Individuals' Agency. *Social Indicators Research*, *159*(3), 1125–1153. https://doi.org/10.1007/s11205-021-02791-8

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. ..*Journal of Statistical Software*, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. Structural Equation Modeling, 14, 464-504.

https://doi.org/10.1080/10705510701301834

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a New General Self-Efficacy Scale. *Organizational Research Methods*, *4*(1), 62–83. https://doi.org/10.1177/109442810141004

Chinman, M. J., & Linney, J. A. (1998). *Toward a Model of Adolescent Empowerment: Theoretical and Empirical Evidence*.

Davidov, E., Muthen, B., & Schmidt, P. (2018). Measurement Invariance in Cross-National Studies: Challenging Traditional Approaches and Evaluating New Ones. *Sociological Methods & Research*, *47*(4), 631–636. https://doi.org/10.1177/0049124118789708

Delgado-Rico, E., Carretero-Dios, H., & Rueh, W. (2012). Content validity evidences in test development: An applied perspective! *International Journal of Clinical and Health Psychology*.

Donald, A., Koolwal, G., Annan, J., Falb, K., & Goldstein, M. (2020). Measuring women's agency. *Feminist Economics*, *26*(3), 200–226.

Duerden, M., Witt, P., Joliff, M., Fernandez, M., & Theriault, D. (2012). Measuring Life Skills: Standardizing the Assessment of Youth Development Indicators. *Journal of Youth Development*, *7*, 99–117. https://doi.org/10.5195/JYD.2012.155

Dupuy, K., Bezu, S., Knudsen, A., Halvorsen, S., Kwauk, C., Braga, A., & Kim, H. (2018). *Life Skills in Non-Formal Contexts for Adolescent Girls in Developing Countries* (CMI Report Number 5.). Center for Universal Education at the Brookings Institution.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Drydyk, J. (2017). Empowerment, agency, and power. In Gender justice and development: Vulnerability and empowerment (pp. 17-30). Routledge.

EASEL. (n.d.). *EASEL Lab*. Retrieved 13 March 2025, from https://easel.gse.harvard.edu/home

Eteläpelto, A., Vähäsantanen, K., Hökkä, P., & Paloniemi, S. (2013). What is agency? Conceptualizing professional agency at work. *Educational Research Review*, *10*, 45–65. https://doi.org/10.1016/j.edurev.2013.05.001

Gai, M. J. P., Cruz, R. M., Viseu, J. N. R., Sales, S. S., & Nunes, C. (2023). Psychometric properties of agency measuring instruments: A systematic review. Ciencias Psicológicas, 17(2), 1–21.

Gandara, F. (2024, April). *Revising Room to Read's Adolescent Life Skills Assessment (ALSA)—Room to Read*. https://www.roomtoread.org/the-latest/revising-room-to-reads-adolescent-life-skills-assessment/

Gandara, F., Sidle, A., & Oulo, B. (2024). *Co-Validation of ALSA and AGAS*.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (Fourth edition). Advances Analytics, LLC.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464. https://doi.org/10.1016/j.labeco.2012.05.014

Heckman, James et al, J., Stixrud, J., & Urzua, S. (2006). *The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior* (Working Paper 12006; NBER Working Paper Series, p. 80). National Bureau of Economic Research.

Hinson, L., Marlow, H., Bhatti, A., Yegon, E., & Izugbara, C. (2021, February). *Measuring Girls' Empowerment*. http://icrw.org/wp-content/uploads/2021/04/ICRW_Measuring_Girls_Empowerment_02.21_Final.pdf

Hitlin, S., & Elder, G. H. (2006). Agency: An Empirical Model of an Abstract Concept. *Advances in Life Course Research*, *11*, 33–67. https://doi.org/10.1016/S1040-2608(06)11002-3

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ibrahim, S., & Alkire, S. (2007). Agency and Empowerment: A Proposal for Internationally Comparable Indicators. *Oxford Development Studies*, *35*(4), 379–403. https://doi.org/10.1080/13600810701701897

Jagosh, J., Macaulay, A. C., Pluye, P., Salsberg, J., Bush, P. L., Henderson, J., Sirett, E., Wong, G., Cargo, M., Herbert, C. P., Seifer, S. D., Green, L. W., & Greenhalgh, T. (2012). Uncovering the Benefits of Participatory Research: Implications of a Realist Review for Health Research and Practice. *The Milbank Quarterly*, *90*(2), 311–346. https://doi.org/10.1111/j.1468-0009.2012.00665.x

Kabeer, N. (1999). Resources, Agency, Achievements: Reflections on the Measurement of Women's Empowerment. *Development and Change*, *30*(3), 435–464. https://doi.org/10.1111/1467-7660.00125

Kabeer, N. (2016). Gender Equality, Economic Growth, and Women's Agency: The "Endless Variety" and "Monotonous Similarity" of Patriarchal Constraints. *Feminist Economics*, *22*(1), 295–321. https://doi.org/10.1080/13545701.2015.1090009

Kehinde, O. J., Dai, S., & French, B. (2022). Item parameter estimations for multidimensional graded response model under complex structures. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.947581

Klein, D. (2018). Implementing a General Framework for Assessing Interrater Agreement in Stata. *The Stata Journal: Promoting Communications on Statistics and Stata*, *18*(4), 871–901. https://doi.org/10.1177/1536867X1801800408

Kwauk, C., & Bragga, A. (2017, November). *Translating Competencies to Empowered Action: A Framework for Linking Girls' Life Skills to Social Change*. The Brookings Institution.

Lautamo, T., Paltamaa, J., Moilanen, J., & Malinen, K. (2021). Psychometric properties of the Assessment Tool for Perceived Agency (ATPA-22) – utility for the rehabilitation of young adults not in education, employment or training (NEETs). *Scandinavian Journal of Occupational Therapy*, *28*(2), 97–109. https://doi.org/10.1080/11038128.2020.1782983

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2022). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. Sociological Methods & Research, 51(4), 1801–1834.

Lerner, R. M. (2009). The positive youth development perspective: Theoretical and empirical bases of a strengths-based approach to adolescent development. In *Oxford handbook of positive psychology, 2nd ed* (pp. 149–163). Oxford University Press.

Lerner, R. M., Theokas, C., & Jelicic, H. (2005). Youth as Active Agents in Their Own Positive Development: A Developmental Systems Perspective. In W. Greve, K. Rothermund, & D. Wentura (Eds.), *The adaptive self: Personal continuity and intentional self-development* (pp. 31–47). Hogrefe & Huber.

Linden, W. J. van der, & Hambleton, R. K. (1997). *Handbook of modern item response theory* (1st ed. 1997.). Springer.

Lloyd, C. B., & Hewett, P. (2009). Educational Inequalities in the Midst of Persistent Poverty: Diversity Across Africa in Educational Outcomes. *Journal of International Development*,

Masur, P. K. (2022). ggmirt: Plotting functions to extend the package “mirt” for IRT analyses (R-package, version 0.0.0.9000). https://github.com/masurp/ggmirt

Mugo, J., Mauro, G., Ngina, P., & Shariff, K. (2022). *The ALiVE Way: Contextualizing the Measurement of Life Skills and Values in Kenya, Tanzania, and Uganda | INEE*. Regional Education Learning Initiative (RELI). https://inee.org/resources/alive-way-contextualizing-measurement-life-skills-and-values-kenya-tanzania-and-uganda

NFER. (2023, November 22). *Life skills assessment development for UNICEF and The World Bank*. NFER. https://www.nfer.ac.uk/international/international-development/our-approach/international-development-case-studies/life-skills-assessment-development-for-unicef-and-the-world-bank/

Pitts, S. C., West, S. G., & Tein, J.-Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. Evaluation and Program Planning, 19(4), 333–350. https://doi.org/10.1016/S0149-7189(96)00027-4

Ogunbiyi, B. O., Bingenheimer, J. B., Baird, S., & Vyas, A. (2025). Measuring adolescent girls' agency. *Journal of Adolescence*, *97*(1), 219–232. https://doi.org/10.1002/jad.12414

Rachel Marcus, Nandini Gupta-Archer, Madeleine Darcy, & Ella Page. (2017). *Girls' clubs, life skills programmes and girls' wellbeing outcomes*. https://tinyurl.com/yc5xnwp8

Raj, A., Dey, A., Rao, N., Yore, J., McDougal, L., Bhan, N., Silverman, J. G., Hay, K., Thomas, E. E., Fotso, J. C., & Lundgren, R. (2024). The EMERGE framework to measure empowerment for health and development. *Social Science & Medicine*, *351*, 116879. https://doi.org/10.1016/j.socscimed.2024.116879

Richardson, R. A. (2018). Measuring Women's Empowerment: A Critical Review of Current Practices and Recommendations for Researchers. *Social Indicators Research*, *137*(2), 539–557. https://doi.org/10.1007/s11205-017-1622-4

Russell, J. (1996). *Agency: Its Role in Mental Development*. Psychology Press.

Ryan, R. M., & Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, *55*, 68–78. https://doi.org/10.1037//0003-066x.55.1.68

Scales, P. C., Benson, P. L., Oesterle, S., Hill, K. G., Hawkins, J. D., & Pashak, T. J. (2015). The dimensions of successful young adult development: A conceptual and measurement framework. *Applied Developmental Science*, *20*(3), 150–174. https://doi.org/10.1080/10888691.2015.1082429

Sen, A. (1999). Development as freedom. New York: Knopf.

Sidle, A. (2019). *Action on Agency: A Theoretical Framework for Defining and Operationalizing Agency in Girls ' Life Skills Programs by Aubryn Allyn Sidle*. https://www.semanticscholar.org/paper/Action-on-Agency-%3A-A-Theoretical-Framework-for-and-Sidle/c852132ee150660471b091440bf1d8f93e868361

Sidle, A. (2020, April). *Measuring Girls' Agency in East Africa—Co-Creating Contextually Specific Tools for Evaluation: Lessons from the AMPLIFY Girls Collective*. Comparative International Education Society (CIES)2020, Miami, Florida.

Sidle, A., & Oulo, B. (2023). Assessment of a Practitioner-Derived Framework for Measuring Girl's Agency in East Africa. *Comparative Education Review*, *67*(2), 331–352. https://doi.org/10.1086/724154

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema*, *35*(3), 217–226. https://doi.org/10.7334/psicothema2022.477

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *1*(26), 100–107. https://doi.org/10.7334/psicothema2013.256

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Sutterlüty, F., & Tisdall, E. K. M. (2019). Agency, autonomy and self-determination: Questioning key concepts of childhood studies. *Global Studies of Childhood*, *9*(3), 183–187.

Tourangeau, R. (1984). Cognitive Science and scale Methods: A Cognitive Perspective. In *Cognitive Aspects of scale Methodology: Building a Bridge between Disciplines* (pp. 73-100.). National Academy Press.

Unterhalter, E. (2019). *Achieving Gender Equality in and through Education*. Global Partnership for Education (GPE). https://www.globalpartnership.org/node/document/download?file=sites/default/files/2019-07-kix-gender-final-english.pdf

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*, *3*(1), 4–69. https://doi.org/10.1177/109442810031002

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.

Yount, K. M., VanderEnde, K. E., Dodell, S., & Cheong, Y. F. (2016). Measurement of Women's Agency in Egypt: A National Validation Study. *Social Indicators Research*, *128*(3), 1171–1192. https://doi.org/10.1007/s11205-015-1074-7

Zimmerman, L. A., Li, M., Moreau, C., Wilopo, S., & Blum, R. (2019). Measuring agency as a dimension of empowerment among young adolescents globally; findings from the Global Early Adolescent Study. *SSM - Population Health*, *8*, 100454. https://doi.org/10.1016/j.ssmph.2019.100454

Ethical Approvals

Ethical approvals for the study were obtained from the Institutional Scientific and Ethics Review Committees at Strathmore University in Kenya (SU-ISERC1529/22) and Makerere University and the Uganda National Council for Science and Technology (SS1726ES), the University of Rwanda (131/CMHS IRB/2023), and the National Commissions for Research and Technology in both Tanzania (2023-152-NA-2022-542) and Malawi (P.12/22/707). All research procedures were performed in accordance with relevant approvals obtained in each country.

Informed Consent

All participants aged 18 and older provided written informed consent. For participants under the age of 18, written informed consent was obtained from a parent or legal guardian, supplemented by written assent from the child.