Room to Read®

Center for Universal Education at BROOKINGS

Life Skills Assessment Technical Workshop
Summary Notes

The Brookings Institution, Washington DC, January 21 – 22, 2020

Thank you for joining us for the Life Skills Assessment Technical Workshop, hosted by Room to Read and the Center for Universal Education at Brookings. We appreciated the lively and candid discussion about this important and complex topic. As noted in the meeting, we plan to organize additional forums in the future to deepen the dialogue and expand the evidence base.

Please find below summary notes from the presentations and discussions. A downloadable zip file of all conference presentations (9 MB) can be accessed here and a list of presentations is provided below in Annex 1.

Assessing and improving validity in life skills measures *(See presentations 1-03 – 1-07 in zip file linked above)*

1. Measurement of life skills needs to be grounded in a thorough understanding of how skills are defined and classified. CASEL and the Harvard EASEL Taxonomy Project have made progress in these areas that others can leverage to ensure measurement constructs are appropriately scoped for the purpose. Too many measurement approaches create scales based on scales which are themselves based on other scales, potentially taking us far from what we would measure if conceptual definitions were our starting point.
2. At the same time, frameworks have largely ignored contextual factors that affect how life skills develop, manifest, and translate into longer-term outcomes. This represents an opportunity for future work.
3. Skill areas of interest may in many cases inherently overlap—we should question whether our assessment is attempting to disentangle constructs that conceptually can't be separated.
4. Unlike skills like math and reading, life skills may not tend to progress monotonically—evidence from California suggest many may decline over the course of adolescence. Relatedly, as beneficiaries age and mature, they may also understand the items differently, which may affect their self-ratings. This has implications for when we measure and how we interpret results and suggests RCTs or other methods using comparison groups may be critical. We may also attempt to design measures that capture the growing sophistication of skills as adolescents age and mature.
5. Validity must be assessed for a particular purpose in a particular context; local factors may substantially affect the way measures work and therefore must also influence our use and interpretation of them.
6. Cognitive interviews using think-aloud and probing methods are a productive approach for improving validity in assessments in that they help to identify misalignment between survey items as intended vs the respondents' understanding of those items.
7. Use of third-party assessments (for example, from teachers and parents) for triangulation to validate students' self-report can be a promising approach, but has challenges—measures must be designed to focus on areas where third-party assessors have knowledge of the student and the construct, and should take into account that these respondents have their biases as well.
8. To reduce the risk of bias in the validation approach itself, consider separating the person or entity doing the validation from the person or entity that developed the original measure.

9. Approaches such as anchoring vignettes, forced choice, and situational judgement tests may help to reduce bias, but can also add complexity to the administration and/or the analysis.
10. Both exploratory and confirmatory factor analyses play an important role and should be integrated into more instrument development processes.

Testing reliability and interpreting results *(See presentations 2-01 – 2-03 in zip file linked above)*

1. Common conventions and rules-of-thumb around Cronbach’s alpha are built around misunderstandings and oversimplifications of the original work. Alpha is influenced not only by reliability but by a range of other factors including the number of items in a construct, sample size, score distributions and the homogeneity of the group.
2. The commonly used 0.70 threshold for reliability was originally proposed as a minimum for the early stages of research only, with higher alphas of 0.90 or greater required to develop good knowledge.
3. Other reliability estimate scores such as KR-20 and Guttman’s lambda’s (particularly lambda-2), have never been as widely used or understood as alpha, but have some advantages that suggest they may be worth exploring.
4. Beyond internal consistency, other categories of reliability such as inter-rater reliability and test-retest reliability may also provide important information about the functioning of an assessment.
5. There may be a tradeoff in time and resources between piloting to improve reliability vs. increasing the sample size to compensate for lower reliability—however the latter strategy may be less likely to be published or regarded externally as acceptable.

Fit-for-purpose assessment: contextualization, measurement design, and interpreting scores *(See presentations 2-04 – 2-06 in zip file linked above)*

1. Contextualization is a particular challenge for life skills measurement and the following elements should be considered:
    a. How skills manifest may vary across geographies, cultures, and age groups, and between relatively privileged and non-privileged groups within a context.
    b. Structural (political and economic) factors affect how life skills translate into ultimate life outcomes.
    c. Local stakeholders’ views on which skills are important varies across contexts.
2. Despite the above challenges, striving for comparability is important for many reasons including justice and equity concerns.
3. We also need to consider the unit of measure: do we need to measure at the level of the individual? The enabling environment? Or both? It is also important to measure the interaction between the two. Assessing the conditions for children to learn certain life skills may in some cases be more appropriate, meaningful, and actionable than assessing learning outcomes at the level of individual children.
4. There is a need to distinguish between standardized assessments and those aligned to a specific curriculum; both may be important depending on the circumstances and purpose. Relatedly, measurement for program evaluation may be very different from measuring to track performance within an education system.
5. We further need to ask to what extent we are measuring *skills* (malleable) vs. *traits* (characteristic, slow to change or may even be somewhat immutable) vs. *moods* (temporary emotional states) and be explicit about our purpose and measures.

6. The time horizon of measurement also needs to be considered: life skills changes themselves may take time to manifest, but also a small change in some life skills could lead to big changes in later outcomes while a big change in some life skills could lead to small changes in outcomes. We also know that during adolescence there can be a high degree of volatility and change in these types of skills and attitudes. How do we take this into account at the time of measurement?

Systems strengthening to integrate and scale assessments *(See presentation 2-07 in zip file linked above)*

1. There is a clear demand by national governments worldwide for effective social and emotional learning interventions and approaches to measuring their effectiveness. How can we respond to this demand?
2. Life skills tends to be thought of as an add-on and relegated to a separate secondary/tertiary subject. Optimizing Assessment for All, in contrast, has designed and piloted measures assessing skills like collaboration and critical thinking as integrated into other subjects like math and social sciences.
3. It is critical to understand each country's starting point with respect to life skills/21st century skills and their expressed policy priorities in order to identify the most effective next steps to move forward.
4. To influence policy, we also need to ask what the threshold for "good enough" measurement may be: the good enough range to make a good decision--to course correct for a program, to invest in a policy reform, etc. Decision-makers need the simplicity; but we are steeped in its complexity. Is this precision functional?
5. In conversations with policymakers, we need to be prepared to resist pressure to turn life skills into a high-stakes test, which may be strong in some contexts.

Some research questions to consider

1. To what extent do life skills contribute to final outcomes in areas we care about, such as the labor market, prosocial/antisocial behavior, family formation, health, and well-being?
2. What contextual factors make up the enabling environment for life skills, and how can measurement of these factors deepen our interpretation of measurement at the individual level?
3. Are certain skills more contextually influenced than others and as a result, do we need to take greater care adapting certain measures across contexts?

Annex 1: Presentations list

Day One

- 1-01: Christine Beggs (Room to Read), opening remarks
- 1-02: David Osher (AIR), slides from introductory panel
- 1-03: Steve Glazerman (IPA), “Assessing and Improving Life Skills Measures: A Research Agenda”
- 1-04: Sonya Temko (EASEL Lab), “Explore SEL: Implications for Measurement”
- 1-05: Allyson Krupar (Save the Children), “Measuring Children’s Social and Emotional Well-being: Validating the International Social and Emotional Learning Assessment”
- 1-06: Ryan Hebert (Room to Read), “Validity in Room to Read’s Life Skills Assessment”
- 1-07: Michel Rousseau (Université du Québec à Trois-Rivières), “Bias in Testing”

Day Two

- 2-01: Michel Rousseau (Université du Québec à Trois-Rivières), “Challenges in Reliability Studies”
- 2-02: Ryan Hebert (Room to Read), “Reliability in Room to Read’s Life Skills Assessment”
- 2-03: Allyson Krupar (Save the Children), “Measuring Children’s Social and Emotional Well-being: ISELA’s Reliability”
- 2-04: Margaret Meagher (AIR), “Fit-for-Purpose Life Skills Assessment: Key Challenges”
- 2-04.5: Wednesday group exercise
- 2-05: Nicole Haberland (Population Council), “Self-Efficacy and Gender Attitude Scales in the Context of GirlsRead! Zambia”
- 2-06: Byrone Wayodi (Asante Africa Foundation), “Creating the Next Generation of Change Agents Today”
- 2-07: Esther Care (Brookings Institution), “Optimizing Assessment for All”

Annex 2: Workshop agenda

Jan 21	Session Description	Speaker(s)
8:30 - 9:00	Coffee and Light Breakfast	
9:00 - 9:35	Agenda Review, Framing Remarks and Participant Introductions	Christine Beggs, Room to Read
9:35 - 10:05	Panel discussion: The state of life skills measurement, relevance to program and policy objectives, key findings of landscape review, priorities to improve the quality of measurement, and the global view with respect to the SDGs.	Chair: Christine Beggs Christina Kwauk, Brookings Institution Esther Care, Brookings Institution David Osher, AIR
10:05-10:45	Q&A and Discussion	Group
10:45 - 11:00	Break	
11:00-12:00	Expert Presentation: Assessing and improving validity in life skills measures - theoretical underpinnings and validation methods.	Chair: Christine Beggs Steve Glazerman, Innovations for Poverty Action (IPA)
12:00-12:30	Q&A and Discussion	Group
12:30-1:15	Lunch	
1:15-2:10	Organizational Presentations: Methods used to assess and improve validity, results of validity testing, assessment adaptation based on validity testing and lessons learned.	Chair: Christine Beggs Sonya Temko, Harvard University Allyson Krupner, Save the Children Ryan Hebert, Room to Read
2:10-2:20	Expert reflections on validity presentations	Esther Care Steve Glazerman Michel Rousseau, Université du Québec à Trois-Rivières
2:20-2:40	Q&A and Discussion	Group
2:40-3:15	Small group work: Utilizing UNICEF India's life skills framework, identify underlying construct assumptions and brainstorm methods for validity testing.	Group
3:15-3:30	Coffee break	
3:30-4:00	Expert Presentation: Bias – Different types of biases, challenges specific to life/socio-emotional skills assessments and strategies to mitigate bias.	Chair: Ryan Hebert Presenter: Michel Rousseau
4:00-4:30	Q&A and Discussion	Group
4:30-4:45	Summary: Key messages, emerging priorities and research questions, Day 2 agenda review and revisions.	Chair: Christine Beggs

Jan 22	Session Description	Speaker(s)
8:30-9:00	Coffee & Light Breakfast	
9:00-9:20	Summary: Day 1 key messages, Day 2 agenda and objectives.	Chair: Christine Beggs
9:20-9:50	Expert Presentation: Expert presentation on reliability including types of reliability and their relevance to life skills/soft skills measurement. Methods for testing reliability, limitations and interpretations. Current prevalent practice and how it should evolve/improve.	Chair: Christine Beggs Michel Rousseau
9:50- 10:15	Q&A and Discussion	
10:15-10:30	Coffee break	
10:30-11:00	Organizational presentations on strategies and experience with reliability testing, including results and adaptations based on results.	Chair: Ryan Hebert Ryan Hebert Allyson Krupner
11:00-11:15	Expert reflections on organizational presentations	Steve Glazerman Esther Care
11:15-12:00	Q&A and Discussion	Group
12:00-1:00	Lunch	
1:00-1:20	Expert Presentation: Fit for purpose: the challenges of adaptation/contextualization, the limits of transferability of skills and measures, interpretation of scores/data, level and focus of measurement.	Chair: Ryan Hebert Margaret Meagher, AIR
1:20-1:50	Small group work: solutions to fit for purpose challenge	Group
1:50-2:25	Organizational presentations on strategies for contextualization and lessons learned, aligning assessment objectives with measurement design, and interpretation of scores/data.	Chair: Ryan Hebert Nicole Haberland, Population Council Byrone Buyu Wayod, Asante Africa
2:25-2:35	Expert Reflection: Mapping assessment/measurement discussion to practice and policy: implications.	Christina Kwauk
2:35-3:00	Expert Presentation: Systems strengthening to scale assessments and integrate into government systems.	Chair: Christine Beggs Esther Care
3:00-3:15	Coffee break	
3:15-3:30	Expert reflection on scaling assessments	Michel Rousseau Margaret Meagher
3:30-4:00	Q&A and Discussion on systems strengthening to scale	Group
4:00-4:30	Summary: Defining a research agenda to improve life/socio-emotional skills measurement, and key workshop messages.	Christine Beggs

Annex 3: List of participants

First Name	Last Name	Organization
Adelle	Pushparatnam	World Bank
Aimee	Reeves	School to School International
Alberto	Begue	UNICEF
Alejandra	De Freitas	FHI 360
Allyson	Krupar	Save the Children
Anne	Mueni Muli	Asante Africa
Byrone	Buyu Wayodi	Asante Africa
Christina	Kwauk	Brookings Institution
Christine	Beggs	Room to Read
Cristobal	Cobo	World Bank
Daniel	Lavan	Education Development Center
David	Osher	American Institutes for Research
Dhiraj	Anand	Room to Read
Diego Luna	Bazaldua	World Bank
Eleanor	Sohnen	Independent
Elizabeth	Kim	International Youth Foundation
Esther	Care	Brookings Institution
Eyerusalem	Tessema	Save the Children
Gemma	Ferguson	Equal Access International
Hajra	Zahid	MasterCard Foundation
Heather	Simpson	Room to Read
Hetal	Thukral	School to School International
Jennifer	Muz	George Washington University
Juliette	Berg	American Institutes for Research
Linda	Tran	Room to Read
Linda	Fogarty	World Bank
Lucina	DiMeco	Room to Read
Manuel	Cardoso	UNICEF
Margaret	Meagher	American Institutes for Research
Masha	Bertling	Harvard University
Meri	Ghorkhmazyan	World Learning
Michel	Rousseau	Université du Québec à Trois-Rivières
Nancy	Taggart	USAID
Nicole	Haberland	Population Council
Nikhit	D'Sa	University of Notre Dame
Nokhanyiso	Mantshongo	Ministry of Education, South Africa
Pamela	Mendoza	Save the Children
Pia	Campbell	International Youth Foundation
Rebecca	Pagel	USAID
Rebecca	Jeudin	Education Development Center
Ryan	Hebert	Room to Read
Scott	Pulizzi	American Institutes for Research
Smita	Das	World Bank
Sonya	Temko	EASEL Lab, Harvard University
Stefany	Thangavelu	Jaurez & Associates
Steve	Glazerman	Innovations for Poverty Action
Victoria	Levin	World Bank
William	Federer	Independent