MEASUREMENT OF HOLISTIC SKILLS IN RCTS: REVIEW AND GUIDELINES

Last updated December 1, 2023

Prepared by:
Karen Macours, Paris School of Economics, LAI Co-Chair
Jessica Williams, J-PAL Policy Associate
Samuel Wolf, Former J-PAL Policy Associate, current MIT Predoctoral Associate

We are deeply grateful to Orazio Attanasio, Jishnu Das, Felipe Barrera-Osorio, Nava Ashraf, Natalie Bau, Pedro Carneiro, and Vladimir Ponczek for their invaluable feedback on the earlier drafts of this report. We are equally appreciative for the insightful comments and suggestions to further improve this resource from Abhijeet Singh, Sule Alan, Asad Islam, Rachid Laajaj, Elise Huillery, Marta Rubio Codina, and Andreas de Barros. We owe a special note of gratitude to all the researchers outside the economics discipline whose comments helped to substantially improve this document: Alex Cristia, Dana Charles McCoy, Pascal Bressoux, Sharon Wolf, Adelle Pushparatnam, and Hiro Yoshikawa. We also thank Celia Hsiao, Nicholas Dowdall, and Eve Hadshar from the LEGO Foundation for their feedback.

For questions, reach out to: lai@povertyactionlab.org

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In order to accurately assess whether an intervention is able to improve a certain skill, that evaluation must be able to measure the skill validly and reliably. This document starts from a review of the measurement strategies used in 237 randomized control trials (RCTs) in the AEA registry, aimed at measuring and improving a wide range of skills beyond typical literacy and numeracy skills in children aged 3-18, including cognitive, social, emotional, creative, and physical skills. Based on this analysis, we propose a set of guiding questions for researchers to ask themselves at the design stage of a new RCT. We pull out key examples from the literature to highlight how others have chosen to answer those questions in their skill measurement work.

The review documents the type of measures that are currently being used, which shows a wide range of self-reported, reports-by others, and observational measures (direct assessment), with a lot of variation spanning from standardized scales adopted from other disciplines, to ad-hoc questions or modules, to lab-in-the-field or game measures specifically designed for the purposes of the RCT. The review further shows a widespread lack of information on the validity or reliability of measures for the study populations in the economics RCT literature: 69% of the studies with papers report no evidence of validity or reliability. In cases where authors borrow a pre-existing tool and establish validity through precedence, rarely does the cited paper match the same context of the RCT sample; less than 0.5% of papers had citations on the key skill measures used which matched the broad geographic contexts of the intervention setting.

The document then introduces various possible approaches to validity and reliability testing and provides examples of studies where thorough validation and innovative, interdisciplinary combination of measurement techniques has been implemented. It emphasizes the need for the research community to publicly share their methodologies and validation attempts to build collective confidence and discusses related challenges and possible solutions.

1. MOTIVATION

J-PAL has recently launched its Learning for All Initiative, which has a core focus on breadth of skills. There is much to learn about whether and how interventions can affect skills that extend beyond literacy and numeracy, including cognitive skills, social skills, emotional skills, creative skills, and physical skills of children.[1] As interest in measuring these outcomes has grown, tackling the measurement challenges for this wider set of skills, which we will refer to as "holistic skills," becomes an important priority.

In order to accurately assess whether an intervention is able to improve a certain skill, that evaluation must be able to measure the skill validly and reliably.[2] Without establishing validity, one can claim effects on an outcome, when in reality, this is not the actual outcome that is being affected. There is a long track record of studies by economists considering literacy and numeracy skills as primary outcomes, and even for such skill measures researchers are sometimes still grappling with measurement challenges. There are far fewer RCTs focusing on the wider set of skills and many of them are very recent.

Many of the measurement questions from the literacy and numeracy literature carry over to measurement of a wider set of skills. Holistic skills can be particularly prone to both random and systematic measurement error for additional reasons, for instance because skills like creativity or perseverance are less tangible and more multidimensional. The risk of invalid measurement leading to erroneous conclusions about the significance, effect size, and overall success of an intervention therefore is important to consider early during the design phase of an RCT.

Targeting the J-PAL research network, composed predominantly of economists, this document aims to provide guidance on the measurement of holistic skills in randomized control trials (RCT)s, starting from a review of current practice in the RCT literature. There are several closely related measurement questions and concerns in the wider literature, including a substantial literature by non-economists as many of the measures economists use were initially developed by other disciplines. This document attempts to flag such links to the wider literature throughout the report, but a more detailed review of these wider questions is outside the scope of this analysis and the guidance proposed.

[1] Parts of the economics literature distinguish instead between non-cognitive and cognitive skills, in which measures of "cognitive skills" can include academic achievement tests which partly capture content knowledge. In this review, "cognitive skills" refers specifically to the more foundational skill set linked to the mental processes of thinking, reasoning or remembering (that can themselves lead to academic achievement). Cognitive skills can be measured, for instance, by tests of problem solving and executive functioning, while numeracy tests would test, for instance, multiplication and division skills.

[2] Validity is "the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests," which is the definition of validity employed in the 2014 version of the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education's "Standards for educational and psychological testing." Reliability is the degree to which an instrument produces the same results under unchanged conditions.

The review was based on a search of the AEA RCT registry, filtering for projects that mentioned "skills" in their registry outcomes.[3] As the language in studies focusing on preschool ages can differ, a search on "preschool" was added to complete the set of studies.[4] All projects that included holistic skills (as defined above) as a primary or secondary outcome variable, that related to education (although interventions are not necessarily themselves in school settings), that overlapped with our age range (3-18 years old), and that measured child outcomes (as opposed to caregiver or teacher outcomes), were included. This yielded a total of 237 studies, upon which subsequent analysis will be focused.

The following sections will first provide a set of guiding questions for researchers to consider at the RCT design stage regarding skill measures, and their validity and reliability. The next section provides background and examples on the types of measures observed in the literature, distinguishing between self-reported measures, those reported by others, and measures based on observation/direct assessments (including games and lab-in-the-field measures). We draw from the review to provide examples, which will primarily be measures of cognitive, social and emotional skills, as those dominate in the literature (see Appendix for details). We then zoom in on the validation approaches used for those measurement tools, pointing to examples of good practices. This is motivated by the findings of the review that showed that there is a lack of evidence on context-specific validity or reliability in most studies, either through referencing previously developed tools or through bespoke measurement development. We conclude with some broader considerations. For more detailed findings including tables, descriptive statistics, and extra analysis, see Appendix 1.

2. QUESTIONS ON SKILL MEASURES TO CONSIDER AT THE RCT DESIGN STAGE

Different RCTs will require different types of methods and investments in skill measurement. Even so, given the numerous measurement challenges related to holistic skills, good practice includes considerations on how to rigorously measure the relevant latent skill traits and how to establish validity and reliability during the design phase. This document therefore starts from a set of key measurement questions we recommend researchers to consider at the RCT planning and proposal stage. To provide researchers with resources and possible approaches to answer those questions, the rest of the document then draws from the existing practice in the literature, providing specific examples and possible guidelines.

[3] While the AEA RCT registry is only one of several social science registries, it is the primary registry used for RCTs with economist co-PIs, i.e. the kinds of evaluations that J-PAL primarily supports. It includes evaluations that do not yet have working or published papers, therefore giving a more comprehensive snapshot of what has ever been tried.

[4] These search criteria will inherently not capture some early childhood parenting program interventions that happen outside the school, which is in line with the focus of this review. For parenting focused RCTs, please see the Global Education Evidence Advisory Panel (GEEAP) 2023 report (Banerjee et al, 2023). Note also that using different but related search terms for the registry yields additional studies. It is unclear if these additional studies would affect the conclusions of this exercise.

Guiding questions:

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
    - Are there existing tools that I could consider using to measure the skill of interest, or do I need to design a new tool or measurement instrument?
    - What are the advantages of each of the different possible measures for capturing the trait of interest? Could I use multiple measures? If so, how will I combine them?
    - Do I have the disciplinary expertise to use or design these measures? Should I collaborate with a co-author from a different discipline?

2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
    - What other outcomes should my measure be correlated with if it truly measures the trait I would like for it to measure, and how will I test this?
    - What can I do to assure that the proposed measure will allow separating the measurement of the latent trait of interest from other factors (e.g. other related traits or some form of response bias)?
    - If I am planning on using a measure someone else designed, has a validation paper been published? Does the context of the validation paper match the context of my evaluation? (Tested on a similar age group, in a similar language, in a similar geography?)

3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
    - What methods can I use to reduce measurement error?

4. How and where will I report on the measurement adaptation, piloting, validity and reliability tests?
    - Should I commit to this reporting in a pre-analysis plan?

5. If my measure is failing some reliability and validity checks, how will I determine if there is an issue with the measure or with the experimental design? For example, may the measure fail to capture the same trait over time or may the experimental variations themselves affect the validity of a measure (e.g. by inducing changes in response patterns/biases) ?

6. How will I adjust my analysis and the write-up of the research results to reflect my findings on the validity and reliability of my measures?

The initial responses to these measurement questions at the design stage, along with considerations of the importance of having valid and reliable measures of a particular skill to answer the research questions in the RCT, can inform decisions about the piloting, adaptation, and final tool selection for the data collection for the RCT.

Given that the answer to the questions will be different for each study, this review does not provide guidance on which outcomes to measure or which tools to use. However, when choosing which skill(s) to measure in their evaluation, researchers are encouraged to thoughtfully consider whether the skill outcomes they aim to measure are indeed the targeted outcomes of the evaluated intervention(s). Researchers are additionally encouraged to assess whether the targeted skill outcomes are likely to have positive returns in the studied context (whether that is in the labor market, for social integration or otherwise). Skill choices for research can unintentionally favor dominant cultural norms, potentially at the cost of alternative cultural expressions, and neglect the holistic development of the child. Skills that are positively valued in some contexts may have negative connotations in others.[5] Culturally important aspects of holistic skills often risk being ignored, further highlighting the need for context-specific diagnostics to determine which skills to measure.[6] Additionally, as labor markets adapt to technological changes, the returns to certain skills may decline, while the returns to others may increase. Ideally, therefore, researchers would avoid assigning favorable or unfavorable status to skills without also providing empirical evidence to back up those associations.

3. SKILL MEASUREMENT METHODS

INTRODUCTION

The tools to measure skills fall into three broad categories: Self-reported (in which students reports their own beliefs or assessment of their skills, often in a survey, without the potential for external verification); Reported by others (in which parents or teachers do the same); and Observed / Direct Assessments (in which researchers directly observe skills through an assessment, game, or other concrete observation). This section provides an overview of current practice on the use of these different methods, as derived from the review, and refers to Appendix 2 for further detail and references on the specific tools.

SELF-REPORTED MEASURES

Self-reported measures were particularly common for skills that were social or emotional in nature. These measures were most frequently used with primary and secondary children, since self-report measures are often not feasible for pre-primary aged children. The skills that were most frequently self-reported, and the measures used to capture them, included: for grit, the Short Grit Scale (Duckworth and Quinn, 2009); for personality, the Big Five Inventory (John et al, 1991 & 2008); for self-esteem Rosenberg's Self Esteem Scale (Rosenberg, 1965); for self-efficacy, the General Self-

[5] For example, certain cultures may highly value politeness and deference to elders, whereas other cultures may place greater value on outspokenness and directness. As discussed in Greenspan and Woodridge's book Capitalism in America, while teaching traditionally valued skills may benefit the proponents of a capitalist system and/or reflect cultural values, it could be at the expense of the child's holistic development. Beyond the labor market motivations, Ann Arnett Ferguson's book Bad Boys: Public Schools in the Making of Black Masculinity underscores the potential pitfalls of promoting "favorable" behaviors, which can inadvertently perpetuate systemic bias and contribute to the school-to-prison pipeline.

[6] For a model of development of an SEL outcome based on culturally specific research in Tanzania, see Jukes et al (2021).

Efficacy Scale (Sherer et al, 1982); and for growth mindset, Dweck's Mindset Instrument (Dweck, 2006). Occasionally, researchers use modified tools combining measures from earlier literature or adding supplemental questions or develop entirely new tools.

Overall, self-reported surveys have frequently been used to measure social skills, self-esteem, and issues related to emotional regulation, and more rarely for measuring creativity, problem solving, or facets of executive function. In addition to using them as direct measures of students' abilities, self-reported measures can also be compared with direct assessments (such as exam scores) to measure students' self assessment of their own abilities (Bobba and Frisancho, 2022).

There are a range of tools used across all of these categories, but relatively few were observed to be used consistently and multi-regionally. Rather, a range of established tools have often been employed to measure highly overlapping latent traits, often without clear resources on how these tools can be differentiated based on context.

MEASURES REPORTED BY OTHERS

Researchers may choose to employ measures that are reported by others either when children are very young, and therefore unable to provide answers for themselves, or when researchers are hoping to get a second perspective from a source that is possibly less susceptible to experimenter demand effects. In the context of education, these measures are most frequently reported by parents or teachers.

Tools relying on reporting by others, are most commonly found in the early childhood development literature: For instance, the Ages & Stages Questionnaire, the Bayley Scales of Infant and Toddler Development, and the Empathy Questionnaire, all of which are designed for children below the age of approximately 7. There is also a set of tools that can be asked to either parents or children, depending on age; these include the Strengths and Difficulties Questionnaire, the Behavior Problems Index, and the Domain-Specific Impulsivity for Children.

When non-established tools were used to capture parent or teacher views, they were often framed around the utility of a second opinion; for instance, one intervention measured parental beliefs on girls' abilities, time-use, and aspirations as a supplement to questionnaires for the girls themselves. Teachers were most often questioned on classroom behaviors, while surveys occasionally questioned employers around young workers' timeliness, attention to detail, and behavior.

OBSERVED/DIRECT ASSESSMENT

A range of techniques have been used to observe skills directly, including through exams/tests, lab-in-the-field games, administrative records, physiological measures, and other forms of direct assessments. Using observed measures and/or direct assessments can help get around many common concerns regarding experimenter demand effects, and can also help circumvent other response biases that can affect self-reported or other-reported measures. Even so, observed outcomes and direct assessments do not automatically provide valid and reliable measures of the latent construct of interest, and for some constructs it can be difficult to obtain (or even to conceive of) relevant observational data that can be collected within a RCT context. For other constructs, tool development may be necessary, and can become part of the contribution of the study. In other cases, tools relying on direct assessment may exist but the costs related to using them may be high.

Among well-established tools, some of the more commonly observed direct assessments were: for cognitive abilities, the Wechsler Intelligence Scales, Raven's Progressive Matrices, the Digit Span test, the Peabody Picture Vocabulary Test, and the Woodcock-Johnson tests; for a range of development skills, the IDELA tests; for self control, the Preschool Self-Regulation Assessment, the Hearts and Flowers task, and the Stroop Test; and for grit, the Alan, Boneva & Ertac Task.

The majority of the tools described can be done as sit-down assessments (for older children) or with test administrators asking the child to perform the given tasks/games (as often done with younger children). Some are slightly more creative in their methodology. For instance, the Lemonade Test assesses empathy by assessing how children react to receiving poorly tasting lemonade that an adult has clearly worked hard on, while the head-toes-knees-shoulders test assesses whether children are able to follow instructions that contrast physical motion with verbal instruction, like being instructed to touch their heads upon hearing the phrase "touch your toes." A well-known test mostly for children younger than the age range in this review are the Bayley Scales of Infant and Toddler Development (often abbreviated as BSID) which include several scales around cognitive, language and motor child development.

Other direct assessments were developed specifically for a given research project by the research teams, following a range of approaches. For instance, researchers developed formal student assessments to capture context-specific "ethnomathematics" skills (Naslund-Hadley et al, 2022), test "market" arithmetic skills (Banerjee et al, 2017) that can go undetected with existing standardized tests, or developed specific games and tasks to measure numerical and spatial abilities (Dillon et al, 2017). Similarly, for socio-emotional skills, there are examples of student evaluations at the intersection between assessments and games developed for specific RCTs, like measures of sensitivity to gaze direction and emotional expressions (Dillon et al, 2017), or attempts to get students to rate the popularity and social characteristics of their peers (Zarate, 2022).

Lab-in-the-field games, often either developed by researchers or adapted slightly from other research projects, were found to measure a range of skills, typically through a set of incentivized tasks. For instance, games focused on grit allowed students to choose between an easy game with a small reward, and a hard game with a larger reward (Alan et al, 2016); one game focused on negotiation tested whether girls could coordinate with their parents to get a larger number of tokens in a manner that required trust and organization (Ashraf et al, 2020), while other negotiation games included risk games, trust games, and public goods games (see Cavatorta et al., 2022). Altruism was measured with dictator games, while time and spatial preferences could be measured through games that assessed preference transitivity between three rewards (a rope, flute, and yo-yo) (Cardim et al, 2022). Lastly, teamwork skills could be measured through games using, for instance, simulated tasks in a team setting (Zarate, pre-analysis plan).

Other direct assessment measures included physiological measures (for an example, see Ye et al, 2022, which uses EEG scans to measure executive function neural responses), administrative records, which largely focus on discipline, and occasionally direct classroom observation. Outside the studies in this review, innovative wearable technologies and corresponding data processing algorithms, such as vests, headbands, microphones, or proximity sensors are being experimentally used to measure a range of metrics tied to holistic skills, including self regulation, language development, and verbal & non-verbal contact between children and caregivers (Lichand et al, 2022 and Romeo et al, 2018). Many of these technology-based tools result in large amounts of observational data and can be coupled with machine learning techniques to obtain indicators of primary outcomes. Another technology is the use of film to record and observe interactions between children and parents or teachers, and coding them afterwards for certain behaviors, without needing to directly put an enumerator in the classroom or home. Film is often used to observe classroom quality and/or teacher behavior (see e.g., Wolf et al, 2018). In an ECD study in Tanzania, however, the researchers filmed the interactions between children and parents and coded for certain child behaviors. Combining such observational data with direct assessment and parent reports can help with adapting well-known measures to a local context (Almås et al, 2023).

COMBINED MEASURES

Combining methods can also help overcome other measurement challenges. One example is Attanasio et al (2019)'s research on early childhood development through play-based learning in Ghana. Researchers chose the well-established IDELA instrument (direct assessment), a comprehensive development and early learning assessment. Because of concerns that the measurement may be affected by "teaching to the test" implications of the play-based curriculum, the IDELA instrument was complemented at endline with a variety of tasks developed at Harvard's Spelke lab, after they had been earlier piloted in the same context. The additional tasks assessed areas similar to those of the IDELA. Combining the standard direct assessments measure with these additional tasks provided a more precise assessment of child development.

4. VALIDITY AND RELIABILITY OF SKILL MEASURES IN RCT STUDIES

INTRODUCTION

Even if the standardized measurement tools may have been validated in many contexts and for the appropriate age groups for some skills used in RCT studies, the review suggests this is more often the exception than the rule. Ensuring the validity and reliability of skill measures to be used in any given new study therefore, more often than not, remains an important task for the researchers aiming to measure impact on skill outcomes in their RCT studies. This is particularly the case as many skills concepts ultimately refer to latent traits that are hard to observe directly and objectively.

The review of the AEA registry revealed, however, that randomized evaluations on holistic skills have a mixed record on validation, and frequently either do not comment on validation or offer only brief descriptions of their validation approach for their measures. This lack of validation can cast doubt on the results of research, particularly when holistic skills are a primary outcome, by questioning whether any observed change in a given measure is actually reflective of a change in the latent trait the research is aiming to measure.

Across the board, 69% of the studies with papers report no evidence of validity or reliability. This holds both for all-economist teams and for teams that included non-economists. It is possible that research teams conducted validity or reliability tests or consulted previous validations when choosing tools but did not include this information in their paper or appendices. Even so, without a discussion of such tests it is hard for the reader to gauge the value of the reported findings on the measured outcomes.

VALIDATION TECHNIQUES EMPLOYED

For the purposes of this review, we divide the empirical evidence that can help assess validity into four categories:

- Face Validity (in which program participants or local experts are asked directly about the interpretation of the items/questions used to construct a given measure)
- Content Validity (which refers to the extent to which a measure assesses a concept in full)
- Construct Validity (which assesses whether a given measure is correlated with other measures attempting to measure the same concept)
- Predictive Validity (which assesses whether a measure is associated with an outcome that, ex-ante, one would expect the concept to correlate with, statically and dynamically).[7]

Similarly important is to consider reliability often tested by analyzing the consistency with which one achieves a given result on a measure (e.g., by evaluating the same person multiple times).

Often validity and/or reliability are investigated during piloting involving a relatively large set of candidates' measurement tools, and the different tests can then be used to remove or hone measures.

Given the relatively rare use of many of these methods, below follows a more detailed discussion of several of these techniques, with the objective of pointing the reader to potential examples to follow in future work.

EXAMPLE OF RELIABILITY CHECKS

Two common calculations for establishing reliability are the test-retest statistics and Cronbach's alpha. These statistical tests can be especially useful as an easy comparison metric when deciding

[7] For tests, these four aspects help evaluate "the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests," which is the definition of validity employed in the 2014 version of the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education's "Standards for educational and psychological testing."

which measures to use in a final analysis. Test-retest reliability is the correlation of the same test taker's results when the test is administered repeatedly. It measures the share of a measure's variance that can be attributed to the underlying trait being assessed. Danon et al (2023) employ test-retest correlations in assessing their battery of instruments around cognitive and social-emotional skills in young adults in Pakistan. After calculating low test-retest correlations for two task-based measurements of grit administered on tablets (much below the 0.7 standard), the researchers decided to drop these assessments, and rely on more reliable measurements in their final analysis.

Cronbach's alpha was perhaps the most frequently employed, especially among interdisciplinary teams.[8] Though Cronbach's Alpha is limited in scope and has received criticism for its misuse (see Sijtsma, 2009), it can remain a useful cursory check of internal consistency that precedes deeper analysis.

EXAMPLES OF INTRA-MEASURE CONTENT AND CONSTRUCT VALIDITY CHECKS

Analysis based on factor analysis is relatively common, though by no means universal. While exploratory factor analysis is often used to determine the number of factors to pull from the data, it can also be used as a validation check on a measurement tool. As described in the online appendix B of Laajaj and Macours (2021), if the factor analysis does not clearly align with the theoretical expected grouping of variables in a certain scale, it brings into question the validity of the interpretation of the scale in accurately capturing what it is supposed to measure. As such, it can help the researchers determine which items in a scale are more likely to represent which underlying latent constructs. This could then point to the need to discard certain items from a scale; shifting an item from one construct to another; or renaming a construct because the pattern of items shows a different latent construct than the one the scale originally intended to measure. As such, factor analysis can help in adapting a measure that has been validated in one context to another context. Factor analyses, both exploratory and confirmatory, were relatively frequently used in our sample. Examples of exploratory factor analysis include Saavedra et al, 2021, who arrived at a set of factors through iterative analysis that removed items with low factor bearings. Iterbeke et al, 2022 used confirmatory factor analysis through Structural Equation Modeling to verify the appropriateness of their factor structures.

Techniques to assess the consistency of answers within a given measurement tool, including Cronbach's alpha and factor analyses can be useful to know whether the tool is consistently measuring a given latent trait. While data collected with a given tool based on multiple questions may score highly via Cronbach's alpha, or separate into a single factor during factor analysis, indicating it is approximating for a latent trait, that doesn't mean the latent trait necessarily is the skill the researcher was aiming to measure. For robust and compelling assessments of validity, combining such approaches with insights from piloting and predictive validity checks, in which the measure is compared to a tangible outcome expected to be correlated with the latent trait being measured, can

[8] The technical elements of calculating Cronbach's alpha are beyond the scope of this review, but on a mathematical level it is equivalent to the expected value of split half reliability.

provide additional confidence on the tools used (see appendix of Laajaj and Macours (2021) for an example considering various skills).

EXAMPLES OF PREDICTIVE VALIDITY CHECKS:

Predictive validity checks, in which measures are compared to observed outcomes (including outcomes of the intervention) that one would expect to be closely associated with the measure, can provide another appealing way to assess validity when done correctly. When skills are measured as outcomes in RCTs, the skill may be directly targeted by the intervention. Or, when skills are not directly targeted but are still expected to be affected by the intervention, a clear theory-of-change that explains the interest in the skill measure is an important starting point. Beyond that, however, spelling out the conceptual reason to believe that that particular skill is going to affect individuals' economic or social outcomes, will help point to outcomes that one expects to be correlated to the skill measures. Establishing predictive validity then implies that the researcher can empirically show associations between the skill measure and those other outcomes. Examples of this approach in the reviewed literature were relatively scarce, but, when conducted, are quite compelling in arguing validity.

One example of this approach was Alan et al, 2018, which used game-like tasks in Turkey to assess grit, the development of which was one of the main focuses of the intervention. Researchers find that baseline task performance predicts academic performance above and beyond what can be predicted by traditional cognitive ability assessments. Though this does not confirm that their tasks measure grit per se (and no validation measure could do so beyond any doubt), it is a compelling data point, because improved academic performance from grit is one of the key ex-ante elements of their theory of change.

Similarly, in Huillery et al, 2022, researchers conducted a range of validity and reliability checks for their holistic skills measures, which attempted to assess conscientiousness, self-control, and grit. For a subset of their measures, particularly teacher assessments, these measures correlated with school behavior, as measured by administrative data, and academic progression over time, both of which are predicted by their theory of change. (For more information on their predictive validity tests, see their companion paper.)

Some tests of construct validity can fit in a similar category of validation, particularly when two very differently measured tests are compared to one another. In theory, one would expect two measures of the same underlying construct (like a survey and a game) to be well correlated with one another if both are valid. Zárate, 2022, for instance, justifies employment of an altruism self-reported scale by noting its historical correlation with peer rating of altruism, completion of altruistic acts like completing an organ donor card, and smiling (some of which straddle the boundary between construct and predictive validity).

One area for attention in looking at the associative or predictive validity of measures is accounting for the dynamic nature of holistic skills in the development of children. When studying holistic skills, there are lots of open questions on whether and how we should expect measures of a given skill at early ages to translate in different measures of that skill or other sills at later ages (e.g., what correlation to expect between a measure developed to capture problem solving in a three-year-old child versus some expected associated problem-solving outcomes later in life).

Another concern that researchers may want to consider is the potential for measures that are cross-sectionally consistent, but time-variant. In this case, particular measures may appear well-calibrated and valid within a baseline cohort but may lose their validity over the course of the intervention; for instance, an intervention may effectively teach students how to respond “correctly” to a given survey rather than improving their underlying latent abilities or preferences. For an example outside the education literature, see Sater et al, 2022, in which an intervention to raise awareness on the pollutant effects of wood burning increased households’ awareness and decreased reported intent to pollute, but not their actual pollution levels as measured by sensors. One explanation of this result is that initially well-calibrated measures became invalid after the intervention, as the intervention may have changed the participants’ interpretation or view on the questions without changing the underlying latent trait. In cases like these, predictive validation with endline data can be particularly helpful.

EXAMPLES OF PILOTING WORK

Piloting can include cognitive (qualitative) interviewing (Peterson et al, 2017) to allow researchers to gain an initial sense of face validity, adaptation to local cultural norms and expressions, possibly followed by quantitative data collection on small samples to undertake initial validity checks.

For instance, Santos et al, 2022 translated the Short Grit Scale into Albanian and Macedonian for a grit-related intervention with 11 to 14 year old students, and held qualitative interviews with students during their pre-intervention pilot to assess Face Validity. While doing so, they noted a limitation that their sample’s age group was younger than the ages for which the Short Grit Scale was originally developed. These actions were taken despite the Short Grit Scale’s widespread usage; researchers expressed that even routinely used tools are best to validate when brought to a new context or language.

Similarly, the pre-analysis plan for Augsburg et al, 2022 commits to piloting all outcome measurement tools, both those that have been previously validated in-context and those tools that have not been. It specifically states that adaptations will be made based on feedback from participants, and surveyors will be required to achieve high inter-rater reliability on measures. The pre-analysis plan also commits to undertake supplementary post-collection validity checks relating to content validity and construct validity through Cronbach’s Alpha, both exploratory and confirmatory factor analysis, and Item Response Theory.

Similar examples appear elsewhere, including researchers removing instruments due to them being too easy or too difficult (see Dean and Jayachandran, 2019), but mentions of such practices (even if they may be commonly used) are overall relatively rare throughout our review. In particular, the studies that involved lab-in-the-field games, as well as studies developing a novel measurement tool, rarely discussed how piloting informed the development and their understanding of validity for their tools, even in appendices. Planning and documenting validation-relevant piloting, including sharing how pilots informed tool development and adaptation, could substantially increase readers' confidence in results, and allow the research community to build on each other's work.

ADDITIONAL APPROACHES:

In addition to the broad validation approaches discussed above, there are a number of quick adjustments that can be helpful in improving measurement.

One of these is acquiescence bias correction. Though projects included in this analysis occasionally mention themes relating to acquiescence bias (in which individuals are likely to answer in the affirmative to survey questions regardless of content), we did not observe many systematic corrections for acquiescence bias. The methodology for doing so, as laid out in the appendix of Laajaj and Macours, 2021, includes calculating the average response difference between non-reverse-coded and reverse-coded items, assigning a resultant acquiescence score, and subtracting it from every non-forward-coded item while adding it to every forward-coded one. This correction may improve the accuracy of later validation approaches like factor analysis and should be considered when researchers believe acquiescence bias to be pertinent.

Another frequently-mentioned threat to validity and reliability is social desirability bias, in which participants may be inclined to provide a socially desirable answer rather than a correct one. Concerns around desirability bias can be a key drawback of survey data, but few studies took efforts to correct it. One that did was Dhar et al, 2022, which measured the social desirability bias of parents through the Crowne-Marlowe module, which elicits attitudes around moderating content to fit the views of others. The study then interacted the social desirability index constructed based on this module with treatment assignment, showing that that social desirability bias was not affecting measurement of these outcome variables more in treatment than in control, and therefore not biasing ITT effets. This practice is now being followed in an increasing number of studies.[9]

SOME ADDITIONAL THOUGHTS ON VALIDITY AND RELIABILITY CHECKS

Across our review, there are a few examples of studies that were limited in their scope or in the tools they felt comfortable using due to a lack of previous robust validation; for instance, McCoy et al, 2021 refrained from attempting to measure social problem solving, empathy, and academic skills in part due to lack of validation in LMICs, while Bøg et al, 2021 mention limited capacity to measure self-efficacy, enjoyment, and motivation as a result of limited validation in Sweden.

For various holistic skills, the most appropriate tool to measure them in a given context or study can remain hard to predict. This suggests that including multiple measure approaches for the same skill can be particularly valuable (when resources allow for it). Cross-validating these measures against one another could yield significant insights, including about possible social desirability bias. Predictive

[9] Ideally the social desirability measure itself should first be validated, in particular as the social desirability measure itself has been found to have low validity in many contexts (Lanz et al, 2022).

validation of multiple measures can also allow researchers to assess their relative validity. For instance, the companion paper to Huillery et al, 2022 found that children's self-assessment of their own conscientiousness, self-control, and grit was more valid than enumerators' scores on observed short behavioral tasks, contrary to the predictions of the vast majority of surveyed experts. As the number of tools to measure holistic skills grow, the possibility for such cross-validation will expand.

When there are multiple measures of the same underlying latent constructs, or scales based on multiple items, researchers in certain cases use factor analysis or Item Response Theory to aggregate the different measures and improve skill measurement. For standardized scales with multiple items, existing algorithms proposed by the publishing owners or authors of a test, often derived in very different contexts and populations, are indeed not necessarily the most efficient way of extracting signal from a series of items. Yet they can offer the advantage of comparability. In certain cases, showing both results with standardized scoring along with any rescoring, and making raw data available for replication, can help the reader evaluate the trade-offs.[10]

When skill-related tools are used in the medical literature, papers often provide detailed information on the process of adaptation and training of the tool, or sometimes refer to companion papers specifically dedicated to the tool validation. Cavallera et al (2023)'s paper on the validation of the Global Scales for Early Development (GSED), for instance, describe the validation design and study sites, justify how the researchers segmented their sample for validation testing, detail the timeline for their data collection, training, and reliability checks, share which statistical checks and cutoffs they used on the data, and describe the high-level process used for developing the item bank and item structure. While the level and scope of this particular data collection go much beyond what is typically feasible in the context of validation of measurement for one RCT, the type of information provided on the validation processes and related analysis in the paper provides a good example RCT researchers could follow on a smaller scale.

5. MEASUREMENT VALIDATION BASED ON THE CITATIONS AND REFERENCES

As researchers build off the literature before them, authors often borrow measurement tools and scales, and sometimes adapt them to their contexts. Instead of providing direct evidence of validity and reliability, they may then reference such earlier work, to increase confidence in their measures.

Reference to other papers for measurement tools, can be broken down to papers providing original or secondary validation of the interpretation of the measurement obtained with a given tool and papers which just used the tool previously (precedent papers). Referencing the validation of the

[10] There is a related debate in the measurement of academic skills, with concerns that even seemingly straightforward measures of academic abilities can yield different effect sizes by arbitrarily weighting test items differently (Bond and Lang (2013). If the effect size varies with the weights assigned, interpretation of results and conclusions can become difficult, in particular when comparability between studies is sought. This is an active debate, with some attempts to provide solutions, such as Chang (2021)'s new Stata command to assess the robustness of an effect to arbitrary scale choices.

interpretation of the tool can provide a greater signal to the reader of its validity and reliability than simply indicating the tool has been used in the past, as frequently used tools are not necessarily the most valid. Referencing validation papers is, however, not yet very common in the RCT literature. Among the subset of papers referring to earlier papers for specific tools, only 37% cite a true validation paper, however, and papers citing validation for age ranges targeted by the intervention and in broadly similar (geographical) contexts are extremely rare (see appendix I for detailed analysis).

Generally, there is also little discussion within the paper itself about the context of any validation paper, how the contexts are similar or different, and how that informs the researchers' choice of the tool. Including this information, however, could greatly assist the reader to evaluate the relevance of any cited validation papers. It would also lead researchers to study the literature carefully to learn if a tool has been validated, or to bring in relevant expertise, e.g., by having a co-author from a discipline that developed relevant skill measurement tools.

For the cases that the tool lacks validation in the relevant contexts, confidence in the results of the RCT will be greatly enhanced if researchers perform their own validation checks, as described in the previous section, and report on them. Similarly, reporting on methods and lessons from any adaptations to tools with prior validation, from altering the structure of the scales to changing the phrasing of the items, will not only increase readers' ability to understand the measurement, but also allow the discipline to assure learning across studies.

6. TAKEAWAYS

Several educational stakeholders are seeking evidence around holistic skill development. Especially in the aftermath of the pandemic, the global education system has not only had to cope with the disruption to academic learning, but also the social isolation and the stunting of children's social, emotional, and cognitive development. While interest in measuring these outcomes has grown, the challenge of measuring latent traits remains.

This resource is not meant to be a prescriptive document on which tools to use when conducting research around holistic skills but aims to provide guidance to researchers on the questions to consider when designing RCT studies with an important focus on skill measurement, to point to the relevance of establishing validity and reliability of the measures for the context of study, and to provide examples of potential approaches to be considered.

Before concluding, it is important to recognize that researchers may not be reporting on validation checks or not conducting them in the first place because of the cost (and possibly even because of potential rejection from publication based on a measure that shows low validity). There are several solutions to respond to these adverse reporting incentives. As a potential solution to the resources required to develop and test skill measures, researchers themselves may want to focus validation analysis on the tools related to the main results. Research organizations, on the other hand, could distribute funding based on proposals' adherence to the guidance in this document, or provide supportive resources such as a measurement database or a validity/reliability reporting template.

Journals could encourage robust holistic skills measurement reporting by committing to publishing a special issue on such topics. J-PAL is committed to exploring solutions to support on this issue. Applicants seeking funding through the Learning for All Initiative explicitly for the design and validation of new tools can apply under the "Pilot Research Projects" category of funding, as long as a direct and credible link with an application of those tools in future RCT work is established.

The review has revealed that there are opportunities to push toward a higher standard of citing, explaining, or conducting original validation of the measurement tools used in RCT studies. While teams may be conducting these sorts of analyses and discussing the validity and contexts of tools behind the scenes, making such thinking public in the body or the appendix of published papers can then serve as a public good for the next researcher facing similar questions in a similar context. Planning to conduct validity checks early in the research process along with robustness checks after data collection can facilitate this process, while also leaving flexibility for the researcher to drop or adjust the measures and transformations proposed in the pre-analysis plan if the authors find the methodology does not properly capture the traits in that specific context. Research teams working across disciplines are probably particularly well placed to come up with better measurements that draw from the strengths of different fields or to design new measures or adapt old tools to new contexts. It is only when readers have confidence in the measurement strategies that they can have confidence in the results, and those results are what can lead to lasting policy changes toward the most impact.

REFERENCES

Alan, Sule, Teodora Boneva, and Seda Ertac. “Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit*.” *The Quarterly Journal of Economics* 134, no. 3 (April 17, 2019). https://doi.org/10.1093/qje/qjz006.

Almås, Ingvild, Orazio Attanasio, and Pamela Jervis. “Economics and Measurement: New Measures to Model Decision Making.” National Bureau of Economic Research, January 1, 2023. https://doi.org/10.3386/w30839.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, Dc: American Educational Research Association, 2014.

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn. “Negotiating a Better Future: How Interpersonal Skills Facilitate Intergenerational Investment.” *The Quarterly Journal of Economics* 135, no. 2 (January 9, 2020). https://doi.org/10.1093/qje/qjz039.

Attanasio, Orazio, Bet Caeyers, Sarah Cattan, Lina Cordona Sosa, Sonya Krutikova, and Peter Leighton. “Improving Early Childhood Development in Rural Ghana through Scalable Community-Run Play Schemes: Programme Impact Evaluation Report.” London, UK: Institute for Fiscal Studies, July 2019. https://ifs.org.uk/sites/default/files/2023-01/Improving-Childhood-in-Ghana.pdf.

Augsburg, Britta, Orazio Pedro Attanasio, Robert Dreibelbis, Edward Nketiah-Amponsah, Angus Phimister, Sharon Wolf, and Sonya Krutikova. “Lively Minds: Improving Health and Development through Play–a Randomised Controlled Trial Evaluation of a Comprehensive ECCE Programme at Scale in Ghana.” *BMJ Open* 12, no. 10 (October 2022): e061571. https://doi.org/10.1136/bmjopen-2022-061571.

Banerjee, Abhijit, Swati Bhattacharjee, Raghabendra Chattopadhyay, Alejandro Ganimian, Rukmini Banerji, Yathish Dhavala, Esther Duflo, et al. “The Untapped Math Skills of Working Children in India: Evidence, Possible Explanations, and Implications ,” 2017. https://www.povertyactionlab.org/initiative-project/street-smart-or-school-smart-leveraging-working-childrens-competencies-teach?lang=fr.

Banerjee, Abhijit, Sylvia Schmelkes, Tahir Andrab, Rukmini Banerji, Susan Dynarski, Rachel Glennerster, Benjamin Piper, et al. “2023 Cost-Effective Approaches to Improve Global Learning - What Does Recent Evidence Tell Us Are ‘Smart Buys’ for Improving Learning in Low- and Middle-Income Countries?” World Bank, 2023. http://documents.worldbank.org/curated/en/099420106132331608/IDU0977f73d7022b1047770980c0c5a14598eef8.

Bobba, Matteo, and Veronica Frisancho. “Self-Perceptions about Academic Achievement: Evidence from Mexico City.” *Journal of Econometrics* 231, no. 1 (September 2020). https://doi.org/10.1016/j.jeconom.2020.06.009.

Bøg, Martin, Jens Dietrichson, and Anna A. Isaksson. “A Multi-Sensory Tutoring Program for Students at Risk of Reading Difficulties: Evidence from a Randomized Field Experiment.” *The Journal of Educational Research* 114, no. 3 (April 24, 2021): 233–51. https://doi.org/10.1080/00220671.2021.1902254

Bond, Timothy N., and Kevin Lang. “The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results.” *Review of Economics and Statistics* 95, no. 5 (December 2013): 1468–79. https://doi.org/10.1162/rest_a_00370.

Cardim, Joana, Leandro Carvalho, Pedro Carneiro, and Damien de Walque. "Early Education, Preferences and Decision-Making Abilities," 2022. https://congress-files.s3.amazonaws.com/2022-08/Early%20Education%2C%20Preferences%20and%20Decision-Making%20AbilitiesEEAupdated.pdf

Cavallera, Vanessa, Gillian Lancaster, Melissa Gladstone, Maureen M Black, Gareth McCray, Ambreen Nizar, Salahuddin Ahmed, et al. "Protocol for Validation of the Global Scales for Early Development (GSED) for Children under 3 Years of Age in Seven Countries." *BMJ Open* 13, no. 1 (2023). https://doi.org/10.1136/bmjopen-2022-062562

Cavatorta, Elisa, Daniel John Zizzo, and Yousef Daoud. "Conflict and Reciprocity: A Study with Palestinian Youths." *Journal of Development Economics* 160 (January 2023): 102989. https://doi.org/10.1016/j.jdeveco.2022.102989

Chang, Andres Yi. "Test Scores' Robustness to Scaling: The Scale_transformation Command." *The Stata Journal* 21, no. 3 (September 2021): 756–71. https://doi.org/10.1177/1536867x211045574

Danon, Alice, Jishnu Das, Andreas de Barros, and Deon Filmer. "Cognitive and Socioemotional Skills in Low-Income Countries: Measurement and Associations with Schooling and Earnings." *Policy Research Working Paper*, February 1, 2023. https://doi.org/10.1596/1813-9450-10309

Dean, Joshua, Seema Jayachandran, Tvisha Nevatia, Sadish Dhakal, Aditya Madhusudan, Akhila Kovvuri, Sachet Bangia, Alejandro Favela, and Ricardo Dahis. "Attending Kindergarten Improves Cognitive but Not Socioemotional Development in India *," 2019. https://economics.yale.edu/sites/default/files/kindergarten_ada-ns.pdf

Dhar, Diva, Tarun Jain, and Seema Jayachandran. "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India." *American Economic Review* 112, no. 3 (March 1, 2022): 899–927. https://doi.org/10.1257/aer.20201112

Dillon, Moira R., Harini Kannan, Joshua T. Dean, Elizabeth S. Spelke, and Esther Duflo. "Cognitive Science in the Field: A Preschool Intervention Durably Enhances Intuitive but Not Formal Mathematics." *Science* 357, no. 6346 (July 6, 2017): 47–55. https://doi.org/10.1126/science.aal4724

Duckworth, Angela Lee, and Patrick D. Quinn. "Development and Validation of the Short Grit Scale (Grit–S)." *Journal of Personality Assessment* 91, no. 2 (February 17, 2009): 166–74. https://doi.org/10.1080/00223890802634290

Dweck, Carol S. 2006. Mindset : The New Psychology of Success—1st ed. New York: Random House.

EASEL Lab, Harvard Graduate School of Education. "Explore SEL." exploresel.gse.harvard.edu, n.d. http://exploresel.gse.harvard.edu/frameworks/62.

Ferguson, Ann Arnett. *BAD BOYS: Public Schools in the Making of Black Masculinity*. S.L.: University of Michigan Press, 2020.

Fernald, Lia, Elizabeth Prado, Patricia Prado, and Abbie Raikes. "Measuring Child Development: A Toolkit for Doing It Right." *World Bank*, June 13, 2018. https://www.worldbank.org/en/programs/sief-trust-fund/publication/a-toolkit-for-measuring-early-child-development-in-low-and-middle-income-countries

Greenspan, Alan, and Adrian Wooldridge. *Capitalism in America: A History*. New York City: Penguin Press, 2018.

Halle, Tamara G., and Kristen E. Darling-Churchill. "Review of Measures of Social and Emotional Development." *Journal of Applied Developmental Psychology* 45, no. 45 (July 2016): 8–18. https://doi.org/10.1016/j.appdev.2016.02.003

Hambleton, Ronald K., and H. Swaminathan. *Item Response Theory: Principles and Applications. Google Books*. Springer Science & Business Media, 2013. https://books.google.com/books/about/Item_Response_Theory.html?id=dUbwCAAAQBAJ

Harter, Susan. "The Perceived Competence Scale for Children." *Child Development* 53, no. 1 (February 1982): 87. https://doi.org/10.2307/1129640

Huillery, Elise, Adrien Bouguen, Axelle Charpentier, Yann Algan, and Coralie Chevallier. "The Role of Mindset in Education : A Large-Scale Field Experiment in Disadvantaged Schools," 2022. https://www.povertyactionlab.org/sites/default/files/research-paper/working-paper_866_Role-of-Mindset-in-Education-Disadvantaged-Middle-Schools_France_Dec2020.pdf

Institute for Child Success. "IMPACT Measures Tool." ecmeasures.instituteforchildsuccess.org. Accessed October 9, 2023. https://ecmeasures.instituteforchildsuccess.org/measures.

Iterbeke, Kaat, and Kristof De Witte. "Helpful or Harmful? The Role of Personality Traits in Student Experiences of the COVID-19 Crisis and School Closure." *Personality and Social Psychology Bulletin* 48, no. 11 (October 20, 2021): 014616722110505. https://doi.org/10.1177/01461672211050515

Jukes, Matthew C. H., Nkanileka Loti Mgonda, Jovina J. Tibenda, Prosper Gabrieli, Grace Jeremiah, Kellie L. Betts, Jason Williams, and Kristen L. Bub. "Building an Assessment of Community-Defined Social-Emotional Competencies from the Ground up in Tanzania." *Child Development* 92, no. 6 (September 13, 2021). https://doi.org/10.1111/cdev.13673

John, O.P., E.M. Donahue, R.L. Kentle, 1991. The Big Five Inventory – Versions 4a and 54. Berkeley CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O.P; L.P. Naumann, and C.J. Soto, 2008. "Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In O.P. John, R.W.robins & L.A. Pervin (Eds.), Handbook of personality: Theory and research, pp 114-158. New York, NY: Guilford Press.

Laajaj, Rachid, and Karen Macours. "Measuring Skills in Developing Countries." *Journal of Human Resources* 56, no. 4 (October, 2021): 1018-9805R1. https://doi.org/10.3368/jhr.56.4.1018-9805r1.

———. "Online Appendices for MEASURING SKILLS in DEVELOPING COUNTRIES List of Appendices APPENDIX A: BRIEF INTRODUCTION to PSYCHOMETRIC CONCEPTS and METHODS USED APPENDIX B: CONSTRUCTION of the INDEXES APPENDIX C: QUESTIONNAIRE DESIGN and SOURCES APPENDIX D: COLOMBIA REPLICATION OTHER APPENDIX TABLES ADDITIONAL APPENDIX REFERENCES," 2019. https://jhr.uwpress.org/highwire/filestream/1591/field_highwire_adjunct_files/0/JHRv56n04_LaajajMacours_OnlineApp.pdf

Lanz, Lukas, Isabel Thielmann, and Fabiola H. Gerpott. "Are Social Desirability Scales Desirable? A Meta-Analytic Test of the Validity of Social Desirability Scales in the Context of Prosocial Behavior." *Journal of Personality* 90, no. 2 (August 16, 2021). https://doi.org/10.1111/jopy.12662

Lichand, Guilherme, Onicio Batista, John Phuka, Roselyn Chipojola, Beverly Laher, Michelle Bosquet Enlow, Anne Elizabeth Sidamon-Eristoff, et al. "The Early Childhood

Development Replication Crisis, and How Wearable Technologies Could Help Overcome It." *Social Science Research Network*, January 1, 2022. https://doi.org/10.2139/ssrn.4162049

McCoy, Dana C., Emily C. Hanno, Vladimir Ponczek, Cristine Pinto, Gabriela Fonseca, and Natália Marchi. "Um Compasso Para Aprender: A Randomized Trial of a Social-Emotional Learning Program in Homicide-Affected Communities in Brazil." *Child Development* 92, no. 5 (May 16, 2021): 1951–68. https://doi.org/10.1111/cdev.13579

Michell, Joel. "Quantitative Science and the Definition of Measurement in Psychology." *British Journal of Psychology* 88, no. 3 (August 1997): 355–83. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Näslund-Hadley, Emma, Juan Manuel Hernández Agramonte, Carmen Albertos, Ana Grigera, Cynthia Hobbs, and Horacio Álvarez Marinelli. "The Effects of Ethnomathematics Education on Student Outcomes: The JADENKÄ Program in the Ngäbe-Buglé Comarca, Panama." *Publications.iadb.org*, April 1, 2022. https://doi.org/10.18235/0004150

Pancorbo, Gina, and Jacob Arie Laros. "Validity Evidence of the Social and Emotional Nationwide Assessment (SENNA 1.0) Inventory." *Paidéia (Ribeirão Preto)* 27, no. 68 (December 2017): 339–47. https://doi.org/10.1590/1982-43272768201712.

Peterson, Christina Hamme, N. Andrew Peterson, and Kristen Gilmore Powell. "Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes." *Measurement and Evaluation in Counseling and Development* 50, no. 4 (October 2, 2017): 217–23. https://doi.org/10.1080/07481756.2017.1339564

Ponczek, Vladimir Pinheiro, and Cristine Pinto. "The Building Blocks of Skill Development." *Bibliotecadigital.fgv.br*, 2017. http://hdl.handle.net/10438/25805

Pushparatnam, Adelle, Jonathan Seiden, and Diego Luna-Bazaldua. "Guiding Questions for Choosing the Right Tools to Measure Early Childhood Outcomes: Why, What, Who, and How," 2022. https://openknowledge.worldbank.org/server/api/core/bitstreams/7150761d-3417-5de7-a7e3-40c1543a7c8d/content

Romeo, Rachel R., Julia A. Leonard, Sydney T. Robinson, Martin R. West, Allyson P. Mackey, Meredith L. Rowe, and John D. E. Gabrieli. "Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated with Language-Related Brain Function." *Psychological Science* 29, no. 5 (February 14, 2018): 700–710. https://doi.org/10.1177/0956797617742725

Rosen, Jeffrey, Elizabeth Glennie, Ben Dalton, Jean Lennon, and Robert Bozick. "Noncognitive Skills in the Classroom: New Perspectives on Educational Research," 2010. https://files.eric.ed.gov/fulltext/ED512833.pdf

Rosenberg, M. 1965. Society and the Adolescent Self-Image. Princeton: Princeton University Press.

Saavedra, Anna, Ying Liu, Shira Haderlein, Marshall Garland, Danial Hoepfner, Kari Lock, and Alyssa Hu. "Knowledge in Action Efficacy Study over Two Years," 2021. https://cesr.usc.edu/sites/default/files/Knowledge%20in%20Action%20Efficacy%20Study_18feb2021_final.pdf

Santos, Indhira, Violeta Petroska-Beska, Pedro Manuel Carneiro, Lauren Eskreis-Winkler, Ana Maria Boudet, Maria Ines Berniell, Christian Krekel, Omar Arias, and Angela Duckworth. "Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students." *SSRN Electronic Journal*, 2022. https://doi.org/10.2139/ssrn.4233803

Sater, Rita, Mathieu Perona, Elise Huillery, and Coralie Chevallier. "The Effectiveness of Personalised versus Generic Information in Changing Behaviour: Evidence from an Indoor Air Quality Experiment *," 2022.

https://www.povertyactionlab.org/sites/default/files/research-paper/Wood%20burning_April%202022.pdf

Save the Children. "Education in Emergencies Toolkit | INEE." inee.org, December 18, 2017. https://inee.org/resources/education-emergencies-toolkit

Sherer, Mark, James E. Maddux, Blaise Mercandante, Steven Prentice-Dunn, Beth Jacobs, and Ronald W. Rogers. "The Self-Efficacy Scale: Construction and Validation." *Psychological Reports* 51, no. 2 (October 1982): 663–71. https://doi.org/10.2466/pr0.1982.51.2.663.

Sijtsma, Klaas. "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha." *Psychometrika* 74, no. 1 (December 11, 2008): 107–20. https://doi.org/10.1007/s11336-008-9101-0

Wolf, Sharon, J. Lawrence Aber, Jere R. Behrman, and Edward Tsinigo. "Experimental Impacts of the 'Quality Preschool for Ghana' Interventions on Teacher Professional Well-Being, Classroom Quality, and Children's School Readiness." *Journal of Research on Educational Effectiveness* 12, no. 1 (October 18, 2018): 10–37. https://doi.org/10.1080/19345747.2018.1517199

Ye, Karen, Anya Samek, Keith Yoder, Jean Decety, Ali Hortacsu, and John List. "Early Childhood Programs Change Test Scores but Do They Change Brain Activity?," 2021. https://www.dropbox.com/s/4b44byyzih1vkoc/Early%20Childhood%20Programs%20Change%20Test%20Scores%20but%20Do%20They%20Change%20Brain%20Activity%3F_Sept2022.pdf?dl=0

Zárate, Román Andrés. "Team Productivity of Immigrants and Natives in Peru – EGAP." egap.org. Accessed October 9, 2023. https://egap.org/project/team-productivity-of-immigrants-and-natives-in-peru/

———. "Uncovering Peer Effects in Social and Academic Skills." *American Economic Journal: Applied Economics* 15, no. 3 (July 1, 2023): 35–79. https://doi.org/10.1257/app.20210583

APPENDIX 1: DETAILED FINDINGS FROM REVIEW OF RCT LITERATURE ON HOLISTIC SKILLS BASED ON AEA REGISTRY

This appendix reports the findings of a review based on a search of the AEA RCT registry, filtering for projects that mentioned "skills" in their registry outcomes. As the language in studies focusing on preschool ages can differ, a search on "preschool" was added to complete the set of studies. All projects that included holistic skills (as defined above) as a primary or secondary outcome variable, that related to education (although interventions are not necessarily themselves in school settings), that overlapped with the age range corresponding to pre-primary to secondary school (3-18 years old), and that measured child outcomes (as opposed to caregiver or teacher outcomes), were included.

This review emphasizes the current state of RCT literature in economics, focusing on the reliability and validity of measures rather than their usage or historical context, and thus distinguishing it from reviews in psychology and reviews of quasi-experimental methods or lab-based interventions (Halle and Darling-Churchill, 2016 and Michell, 1997).

1.1 BROAD CHARACTERIZATION OF STUDIES COVERED BY THE REVIEW

The search yielded a total of 237 studies. Nearly all of these studies were RCTs, given their presence in the registry, but they differed on many other aspects. 122 had publicly available working or published papers, while 115 did not. Of the 122 studies with papers, 25 contained a separately attached appendix, which we analyzed as well. While almost all projects had at least one economist as PI (principal investigator), 98 projects (41%) had interdisciplinary PI teams, including education researchers (38 projects), psychologists (30), sociologists (14), public policy researchers (12), and health researchers (10). 149 projects (63%) were based in low- and middle-income countries. Represented regions included Europe (52), Latin America and the Caribbean (46), South Asia (41), Sub-Saharan Africa (30), and North America (29), with some smaller representation in the Middle East and North Africa (8), East Asia (8), and Southeast Asia (8). Lastly, 132 projects (56%) included a J-PAL affiliated researcher. The vast majority of projects (86%) concluded fieldwork between 2013 and 2023.

Of the 237 studies evaluated in this review, a large number of distinct interventions were observed. 139 interventions (59%) were school-based, 94 (40%) were based at locations outside of the school building such as in homes or in a virtual space, and 5 (2%) locations were unclear from publicly available information. We further outline a few common non-mutually exclusive categories of interventions based on keyword searches, though each of these is likely an undercount, and is intended more to give an overview of the broad range of interventions examined.

The first group of interventions revolved around changes in the classroom. Per our calculations, 29 studies involved curriculum changes, while 17 involved tutoring and 14 related to pedagogy. Another group of interventions targeted inputs that occur behind the scenes of live teacher-student interactions: 21 related to teacher training, and 8 involved education-related technology (Ed tech). A third intervention category emerged around young children. Early childhood education was identified as a component in at least 22 studies, preschool was mentioned in 15, and 11 interventions involved learning through play. Additionally, there were several studies that went beyond students and lesson plans: 8 involved cash transfers, 5 involved incentives, and 26 were at least partially parent-focused in targeting. Studies related to older learners, including from secondary school, also focused on vocational training (15) and entrepreneurship development (4).

Studies targeted a range of age groups, categorized for this review by broad education/development level: pre-primary age (up to age 5), primary school age (ages 5-11), secondary school age (ages 12-18), and post-secondary age (ages 18+).[11] Target populations sometimes blend over multiple age groups, whereas these categorizations rely on where the majority of the target population range fell. Interventions focused on the secondary school age range made up the largest share (33% of all studies), and almost tied for second most common were interventions that target primary (24%) and those that target preschool (22%) ages. 8% targeted primary and secondary ages students equally and 5% of interventions targeted the post-secondary age group primarily, while another 8% were unclear.

1.2 SKILLS MEASURED

The review focuses on the broad range of skills (learned abilities) that go beyond literacy and numeracy (the learning outcomes that are most often measured in schools). We refer to these skills as "holistic skills," but arguments and findings apply to other commonly used terminology, including socioemotional skills, soft skills, life skills, 21st-century skills, cognitive and noncognitive skills, non-academic skills, and many more. In early childhood, cognitive skills are often measured through early language and math abilities, and while some researchers refer to these as "pre-literacy" and "pre-numeracy" skills, they are included here as holistic skills, rather than purely academic skills, exactly because they are often seen as proxying for cognitive outcomes. Furthermore, researchers sometimes define skills quite broadly, including outcomes such as attitudes, aspirations, behaviors, or mental health, even if these outcomes may not necessarily be acquired (learned) abilities. This review purposely does not take a stand on whether such outcomes should be considered as skill measures, and instead follows the classification used by the authors of the different studies, while flagging the need to carefully consider this in future studies.

Of the 237 RCTs, 53% measured both holistic skills and general academic achievement skills, and over half of those defined both skill sets as primary outcomes. Only 16% of RCTs measured academic achievement as a primary outcome and holistic skills as a secondary outcome. About 7% measured holistic first and academic achievement secondarily, and 43% measured only holistic skills. These findings may however be an artifact of the search terms used.

To provide more insights on the types of holistic skills that are being measured, to the degree possible, we distinguish between emotional, social, cognitive, physical, and creative skills.[12] There are

[11] These are admittedly broad age categories, and how skills are expressed within these categories can vary widely. For example, measures for early cognitive skills are typically different for 0–3-year-olds versus 3–5-year-olds. As each study in the review would include measures for cohorts relevant for the specific intervention that is being evaluated, however, RCTs on preschool interventions may include children younger than 3 or older than 5. The age-categorization therefore uses the age category in which the majority of the age range fell, aggregates all children up to age 5 into one pre-primary category.

[12] This particular skill categorization is taken from EASEL Lab, "LEGO's Skills for Holistic Development," Explore SEL, 2019, http://exploresel.gse.harvard.edu/frameworks/62/terms, and uses the following definitions:

many other ways skills can be categorized and we refer to ExploreSEL to see how the categorization used here maps into other possible categorizations.[13]

Furthermore, one outcome can fall into multiple categories. For example, problem solving can be counted as a cognitive and creative skill. In addition, sometimes studies measured several outcomes which covered multiple skill categories. Therefore, the percentages presented below do not sum to 100.

Skill Classification: % of total within category	All	(1) Studies with paper	(2) Studies without paper	(3) Paper + inter-disciplinary	(4) Paper + no inter-disciplinary	(5) All + inter-disciplinary	(6) All + no inter-disciplinary	(7) Paper + LMIC	(8) Paper + HIC	(9) All +LMIC	(10) All+HIC
Social	44	46	41	40	49	45	42	51	38	45	40
Emotional	58	56	60	54	56	61	54	55	56	58	56
Cognitive	38	40	35	48	34	40	36	39	40	36	40
Creative	7	7	8	6	7	4	9	4	8	6	8
Physical	8	8	7	10	7	7	8	14(**)	0(**)	11(**)	2(**)
Unclear	16	13	18	16	10	14	17	12	13	18	11
Social & Emotional	19	19	20	14	21	22	17	18	21	17	22
Social & Cognitive	4	5	4	6	4	4	4	5	4	4	5
Emotional & Cognitive	6	7	5	6	7	7	6	3(**)	15(**)	5	9
Social & Emotional & Cognitive	12	11	13	16	8	13	12	16(**)	4(**)	15(*)	7(*)

Stars indicate significance level of difference between even and uneven columns with * (10%); ** (5%) significant levels.

Of all the entries considered for the review, note that emotional (58%), social (44%) and cognitive skills (38%) are the top three out of the five buckets, with much fewer studies measuring creative (7%) and physical (8%) skills.[14] The most common combination was the social and emotional skills pair (19% of all studies), but studies measuring outcomes across all three (emotional, social, and cognitive) were similarly common (12% of all studies). In general, skills measured in studies in LMIC

Emotional skills are the abilities to "understand, manage and express emotions by building self-awareness and handling impulses, as well as staying motivated and confident in the face of difficulties". Cognitive skills encompass the "concentration, problem solving and flexible thinking by learning to tackle complex tasks and building effective strategies to identify solutions". Social skills are the abilities to "collaborate, communicate and understand other people's perspectives through sharing ideas, negotiating rules and building empathy". Creative skills encompass "Coming up with ideas, expressing them and transforming them into reality by creating associations, symbolizing and representing ideas and providing meaningful experiences for others". Physical skills are defined as "being physically active, understanding movement and space through practicing sensory-motor skills, developing spatial understanding and nurturing an active and healthy body". Note that physical skills are different from physical health outcomes, such as height-for-age measures, which are common in child development research but outside the scope of this review.

[13] Following the mapping, tests of executive function were classified as measuring cognitive skills, while emotional self-regulation would fall under emotional skills.

[14] Of the 8% of the evaluations that measured physical skills, 83% measured outcomes for children below 5 years old.

and HIC countries are remarkably similar, though papers based in LMICs were significantly more likely to measure the three-way combination as well as physical skills compared to papers based in HICs.

1.3 MEASUREMENT TOOLS

Within these broad categories, 115 RCTs (49% of all studies) include self-reported measures, 51 studies (22%) include measures reported by others, and 95 studies (40%) include observational measures/direct assessments. Note that numbers don't add up to 100%, as many studies utilized multiple measures. A few studies with scarce details on the AEA RCT registry didn't have any measures that we could clearly classify on one of the categories.

Self-reported measures generally consist of student surveys, which can measure a host of relevant skills. Likewise, measures "reported by others" are most frequently surveys or reports by teachers or parents. Observational methods/direct assessments are more varied. They could include examinations or tests (measured in 23% of all studies), which includes established tests (like the Raven or Wechsler Intelligence Scales), or ad hoc examinations or games designed by the researchers for the particular study. Observational methods further include lab-in-the-field games (8%),[15] administrative records (5%), physiological measures (1%), and classroom observations (1%).[16] In general, these broad categories did not vary all that much by discipline nor by income level of the country, with some notable exceptions. Teams with interdisciplinary teams were less likely to use lab-in-the-field games. RCTs in high-income countries were more likely to use administrative records than those evaluations in low-income countries.

Type of measurement tools (%)	All	(1) Studies with paper	(2) Studies without paper	(3) Paper + inter-disciplinary	(4) Paper + no inter-disciplinary	(5) All + inter-disciplinary	(6) All + no inter-disciplinary	(7) Paper + LMIC	(8) Paper + HIC	(9) All +LMIC	(10) All+HIC
Student survey	49	57 (**)	39	54	59	48	48	59	54	46	52
Student exam/ test/ task	23	26	19	26	25	23	22	26	25	21	25
Lab-in-the-field games	8	11	6	2 (**)	17 (**)	4(**)	12(**)	15 (**)	4 (**)	10	5
Physiological measures	1	2	1	4 (**)	0 (**)	2	1	0 (**)	4 (**)	0(**)	3(**)
Parental survey	8	9	8	6	11	7	9	11	6	9	8
Teacher report	3	3	1	4	3	2	2	1	6	1(*)	5(*)
Classroom observation	1	2	1	2	1	1	1	1	2	1	2
Administrative records	5	4	6	2	6	6	4	3	6	2(**)	10(**)

Stars indicate significance level of difference between even and uneven columns with * (10%); ** (5%) significant levels.

In some cases, researchers generated their own tools, or used tools that they pulled from a specific recent paper. However, in other instances, researchers used tools that were well established in the

[15] Lab-in-the-field games refers to those games (often incentivized) that use some controlled elements that would typically be found in traditional lab experiments, but that take place in natural, real-world settings.

[16] Physiological measures can include EEG scans, heart rate monitors or blood pressure gauges, which are sometimes used to measure emotional-skills-adjacent outcomes such as stress. The 3 studies included in this review that used a physiological measure all used EEG scans. See the end of the subsection "direct assessments."

literature (though not necessarily validated in the study context), with clear names and a history of use. This made up 48% of all projects evaluated, and 63% of projects with papers. Of studies with papers, we calculated 61% of measures as being entirely borrowed from previous research, 14% as being entirely new, and 23% as being a combination of new and borrowed.

Appendix 2 lists 36 distinct well-established tools that were used in multiple studies covered by the review to measure holistic skills outcomes (though this is likely a lower bound given that it is not always possible to identify which tools were used from the available information). Of these, 14 were entirely self-reported, 6 were entirely reported by others, 2 were a combination of the two, and 19 were observational.

For other sources on potential tools for skill measurement, see, for instance, the INEE Education in Emergencies toolkit (Save the Children, 2017), the IMPACT Measures repository from the Institute for Child Success, The Play Accelerator Research: Tools Analysis prepared by RTI (Henny et al, 2020), the SIEF/World Bank ECD Measurement Inventory (Fernald et al, 2017) along with an accompanying resource "Guiding Questions for Choosing the Right Tools to Measure Early Childhood Outcomes" (Pushparatnam et al, 2022) or the Rosen et al (2010) booklet for Noncognitive Skills in the Classroom: New Perspectives on Educational Research.

1.4 VALIDATION TECHNIQUES EMPLOYED

Through the 95 papers with an appendix, validation was relatively scarce: 13% referenced content validity checks, 24% discussed construct validity checks, 9% predictive validity checks, with another 12% discussing piloting-related validity checks across the categories above. (Note that we did not quantify the evaluation of face validity, due to the lack of direct discussion related to face validity, and difficulty of judging it independently, in the majority of papers.) The review shows that interdisciplinary teams were more likely to report a content validity check than an entirely economist team.

Usage of validation tools: % of total within category	All	(1) Studies with paper + appendix	(2) Papers without appendix	(3) Paper (app) + inter disciplinary	(4) Paper (app) + no inter disciplinary	(5) Paper (app) + LMIC	(6) Paper (app) + HIC
Report potential threat to validity from tool	17	26	11	28	25	22	34
Report validation-relevant piloting	6	12	2	16	10	17	3
Report reliability check	4	3	1	1	2	1	0
Report content validity check	7	13	3	22(**)	8(**)	12	14
Report construct validity check	11	24	3	31	21	20	31
Report predictive validity check	4	9	0	6	11	10	9
Report no validation check	69	45	85	47	44	43	49
Factor analysis	7	16	1	19	14	15	17
Exploratory factor analysis	3	7	1	6	8	3	14

Confirmatory factor analysis	3	5	1	6	5	5	6
Principal component analysis	3	7	0	9	6	5	11
Test-retest Reliability	1	2	1	3	2	3	0
Cronbach's Alpha	8	17	2	25(**)	13(**)	13	23
Intraclass correlation test	0	1	0	0	2	2	0
Acquiescence bias correction	0	1	0	3	0	2	0
Item response theory	3	4	2	3	5	7	0

Stars indicate significance level of difference between even and uneven columns with * (10%); ** (5%) significant levels.

Within these categories, particularly commonly used validation approaches included Cronbach's Alpha (17%) and factor analysis (16%), with both confirmatory and exploratory factor analysis used. The review found that interdisciplinary teams were more likely to report a Cronbach's alpha than entirely economist teams. Test-retest reliability measures were reported with considerably less frequency in the reviewed studies compared to other validity/reliability checks. Only a few studies used other approaches to improve measurements through acquiescence bias corrections or item response theory.[17] Instances of predictive validity checks, or detailed descriptions of piloting work are rare.

1.5 REFERENCING VALIDATION AND PRECEDENT PAPERS

To increase the confidence in the measurement tools used, researchers often cite either validation papers or precedent papers. "Cited validation papers" would be papers that perform validation checks for a tool, as determined by the title and sometimes the abstract of the work cited. Sometimes this category overlaps with tool development papers, as long as those development papers are clearly performing validation checks (Harter, 1982). Precedent papers are those works cited in explanations of the measurement tool which do not clearly demonstrate from the title or abstract that the paper's main focus was on the validation.

While 63% of papers cite a precedent paper, only 37% of papers cite a true validation paper when laying out their measurement strategy. The sample of papers which cite validation papers was so low, that we could not conclude any significant variation in these results between interdisciplinary and economist-only teams, nor between HIC and LMIC evaluations.

When establishing measurement credibility, it is arguably important not just that the interpretation of scores obtained with borrowed tools are validated, but also that the interpretation of the scores is validated in a context that matches the context of the study. For every study that cited a validation paper, we therefore checked whether the validation paper sample matched the age range, the global regional geographic context, and the country context of the study. These are not the only contextual

[17] Acquiescence bias refers to the tendency of an individual to systematically agree (yea-saying) or disagree (nay-saying) with questionnaire items, regardless of their content. Item Response Theory offers a structural way of using a set of items to measure a latent ability or trait. It is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters. For a general introduction to IRT, see Hambleton and Swaminathan (2013).

considerations. For example, a researcher may also want to know if an established tool was tested to be used in a school or home setting to match their intervention setting. Sometimes, studies would cite multiple validation papers, and in these scenarios, we looked at whether all cited validation papers matched the context or not. For example, Ponczek and Pinto's 2017 paper on the Building Blocks program in Brazil cites a validation paper (Pancorbo & Laros, 2017) for the Social and Emotional or Non-cognitive Nationwide Assessment (SENNA) instrument that was also validated in Brazil. This would be an example of a tool validation matching not only the regional context, but even more precisely, the country context.

Only 18 papers (15%) cited a validation study which matched the age range in the current intervention (defined as at least half of the age range of the intervention overlapped with the validation). A perfect geographic context match with all the cited validation papers within a study were even less frequent. Based on our observations of citations, 11 papers (9%) had all their tools match the regional contexts of their validations and only 5 papers (4%) used tools with validation in the same country as their intervention country. Generally, there is also little discussion within the paper itself about the context of any validation paper, how the contexts are similar or different, and how that informs the researchers' choice of the tool.

1.6 RESULTS AND IMPACT DIRECTIONS

A typical literature review and research agenda would provide insights into synthesizing the results that these evaluations have on their target population. As the main focus of this review has been on measurement methods, the objective here is, however, different. Additionally, given our broad definition of holistic skills, we have captured a wide range of interventions which measure very different outcomes in children, making it difficult to compare and draw generalizations across the interventions' impacts and across a wide age range. Finally, as described in the previous sections, many studies do not share details about the validity and reliability of the measurement tools used, thereby calling into question whether the outcomes they observe are truly reflecting the latent traits they sought to measure. That said, it is arguably informative to summarize whether studies reported a positive, negative, mixed, or null effect on the holistic skills development of the child.

Results direction (% of total within category)	All Studies with paper + appendix	Paper + Validation Reported	Paper + No Validation Reported
Positive	49	51	47
Negative	6	6	7
Mixed	19	21	17
Null	25	21	30

While there is no stark difference in the direction of results between papers reporting validation techniques versus those that do not, the probability of a null result was found to be higher in papers without validation. A possible interpretation for this finding could be that higher random measurement error in papers without validation leads to more attenuation bias, hence possibly erroneously concluding there is no impact on a given skill while there is.[18]

APPENDIX 2: MEASURES OBSERVED IN THE REVIEW

APPENDIX 3: PAPERS OBSERVED IN THE REVIEW

[18] Of course, lack of validation may also lead to unawareness about non-random measurement error, which will lead to different biases.