



Creating Short Forms of Early Childhood Development Measures: A Framework for Quantifying Statistical, Conceptual, and Practical Tradeoffs in Direct Assessment

Jonathan Seiden
Harvard University

Direct assessments of early childhood development (ECD) are a cornerstone of research in developmental psychology and are increasingly used to evaluate programs and policies in lower- and middle-income countries. Despite strong psychometric properties, these assessments are too expensive and time consuming for use in large-scale monitoring or national-level surveys. Short forms of direct assessments can provide some benefits of direct assessment at substantially lower cost and complexity. However, selecting the best items for inclusion on shorter forms is not a straightforward task. Traditional approaches to creating short forms, which rely on statistical properties of items, can neglect important non-statistical considerations and result in narrowed construct coverage that does not maximize improvements in usability. Automated Test Assembly (ATA) is an ideal approach to generate optimal forms given numerical constraints, but can be difficult to operationalize. This paper proposes a theoretical framework for an empirically driven, human-centered process to create short forms of ECD direct assessments. It builds on the goals of ATA in an accessible manner by evaluating items across three dimensions: statistical (how reliably items distinguish between children with higher and lower development), conceptual (how representative items are of the constructs being assessed), and practical (how time-consuming and difficult items are to administer). Having defined this framework, the paper then applies it to the International Development and Early Learning Assessment, a popular direct assessment of ECD, to suggest a general-purpose short form selected after considering these three dimensions.

VERSION: November 2025

Suggested citation: Seiden, Jonathan. (2025). Creating Short Forms of Early Childhood Development Measures: A Framework for Quantifying Statistical, Conceptual, and Practical Tradeoffs in Direct Assessment. (EdWorkingPaper: 25 -1143). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/cld3-kw12>

**Creating Short Forms of Early Childhood Development Measures: A Framework for
Quantifying Statistical, Conceptual, and Practical Tradeoffs in Direct Assessment**

Jonathan Seiden

Harvard Graduate School of Education

Acknowledgements: Sincere thanks to Save the Children US for access to data and the extensive inspiration and guidance from Dr. Lauren Pisani. This paper benefited from several rounds of constructive feedback from Professor Andrew Ho and members of the Harvard Graduate School of Education Measurement Lab as well as Professor Dana McCoy and members of the Settings for Early Education and Development Lab. I greatly appreciate the valuable feedback from subject matter experts that shared their expertise and desires for a short form IDELA. Finally, this work would not have been possible without the tens of thousands of children worldwide who participated in the studies that collected IDELA data, their caregivers who consented to their participation, and the hundreds of data collectors who administered the assessments.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Abstract

Direct assessments of early childhood development (ECD) are a cornerstone of research in developmental psychology and are increasingly used to evaluate programs and policies in lower- and middle-income countries. Despite strong psychometric properties, these assessments are too expensive and time consuming for use in large-scale monitoring or national-level surveys. Short forms of direct assessments can provide some benefits of direct assessment at substantially lower cost and complexity. However, selecting the best items for inclusion on shorter forms is not a straightforward task. Traditional approaches to creating short forms, which rely on statistical properties of items, can neglect important non-statistical considerations and result in narrowed construct coverage that does not maximize improvements in usability. Automated Test Assembly (ATA) is an ideal approach to generate optimal forms given numerical constraints, but can be difficult to operationalize. This paper proposes a theoretical framework for an empirically driven, human-centered process to create short forms of ECD direct assessments. It builds on the goals of ATA in an accessible manner by evaluating items across three dimensions: *statistical* (how reliably items distinguish between children with higher and lower development), *conceptual* (how representative items are of the constructs being assessed), and *practical* (how time-consuming and difficult items are to administer). Having defined this framework, the paper then applies it to the International Development and Early Learning Assessment, a popular direct assessment of ECD, to suggest a general-purpose short form selected after considering these three dimensions.

Keywords: direct assessment, early childhood development, measurement, psychometrics, short form, International Development and Early Learning Assessment

2. Introduction

Over the past 20 years, early childhood has been elevated as an important time for human development with early interventions yielding life-long effects (Black et al., 2017; Grantham-McGregor et al., 2007). As a result, many low- and middle-income countries (LMICs) have dramatically increased investments into programs and policies designed to promote early childhood development (ECD) (Jeong et al., 2021; Khatib et al., 2020; McCoy, Salhi, et al., 2018). Alongside increased investment in this area of policy came demand for ECD assessments that can reliably measure the impacts of policies and programs, generate evidence of existing gaps, and monitor progress over time. Researchers have responded by developing several high-quality direct assessments of ECD specifically for use in diverse international contexts, substantially lowering the cost and complexity of collecting data in LMICs (Fernald et al., 2017).

Direct assessment is typically viewed as a high-quality modality to accurately measure ECD (Fernald & Pitchik, 2019; Munoz-Chereau et al., 2021). However, even with instruments tailored for use in low-resource settings, direct assessment is expensive and time-consuming and thus rarely used for large-scale data collection. Short forms could enable researchers to assess ECD at lower cost and complexity while retaining the benefits of direct assessment. A traditional approach to creating short forms in psychometric work focuses on the statistical properties of items and selecting those with the highest reliability, but this can undermine the validity of scores and ignores the practical considerations of assessment (American Educational Research Association et al., 2014; John & Soto, 2007; Little et al., 1999; Smith et al., 2000; Widaman et al., 2011). Automated Test Assembly is a more advanced approach that can balance additional criteria and produce optimal forms given numeric constraints, but can be difficult to operationalize given the complexity of programming and requires users to create hard constraints for the desired form (Linden, 2005).

The goals of this paper are twofold: 1) to introduce a multi-dimensional human-centered framework to aid in the creation of short forms of ECD that attempts to draw on the strengths of Automated Test Assembly at lower complexity and with more flexibility, and 2) to demonstrate the utility

of this framework by applying it to the International Development and Early Learning Assessment (IDELA) to propose a balanced short form. This framework assesses the items on an assessment according to *statistical*, *conceptual*, and *practical* dimensions. In doing so, tool developers can use information on these three dimensions to create a short form that simultaneously maximizes reliability, reduces complexity, and preserves the breadth of the construct being studied. As an open-source direct assessment of ECD for children aged 3.5-6 years old used for research and evaluation in over 75 countries, the IDELA represents an ideal candidate for a short form that could extend the benefits of direct assessment in programmatic monitoring or large scale data collections to situations where the full assessment would not be viable (Save the Children, 2019, 2024).

To apply this framework, I assemble data to quantify the statistical, practical, and conceptual properties of the IDELA assessment. It then combines information across dimensions to visualize the tradeoffs of including various subtasks on a short form IDELA and applies human judgement to select eight subtasks for a balanced short form. After creating a balanced short form, I compare its psychometric properties against the full assessment and a short form using statistical criteria alone to select subtasks.

3. Measurement of ECD in Low- and Middle-Income Countries

The systematic measurement and quantification of ECD, as many constructs in psychology, has historical roots in high-income countries and with samples of children more advantaged than the general population (Gladstone et al., 2010; Henrich et al., 2010). Since early childhood has been prioritized on the global development agenda, investments in promoting ECD in LMICs has increased demand for the measurement of ECD outcomes and required rethinking how ECD is measured (Black et al., 2017; Fernald & Pitchik, 2019; Grantham-McGregor et al., 2007; Jeong et al., 2021). Well-studied measures of ECD from High-Income countries have been successfully adapted for use in other linguistic and cultural contexts, but typically require highly qualified enumerators and specific standardized materials, resulting in prohibitive costs for regular use in low-resource settings (Rubio-Codina et al., 2016; Sabanathan et al., 2015). Even when “gold-standard” tools of ECD such as the Bayley Scales of Infant and Toddler Development have been successfully contextualized, their norms are often invalid in LMICs and the

complexity of their administration make widespread use in LMICs impractical (Cromwell et al., 2014; Madaan et al., 2021; Pendergast et al., 2018; Pisani et al., 2015).

This situation created demand for relevant and practical tools for use in LMICs across diverse cultural and linguistic settings (Fernald et al., 2017). Since the mid-2000s, there has been a dramatic increase in the creation, availability, and usage of ECD measurement tools designed for broad use in LMIC settings, as well as some for specific regional, cultural, and linguistic areas (Fernald et al., 2017). Notable early examples developed specifically for use in LMICs were the Malawi Developmental Assessment Test (Gladstone et al., 2010) and the East-Asia and Pacific Scales for Early Development (Rao et al., 2014). Since then, the field has expanded and researchers seeking to measure early childhood development in LMICs today face a dizzying array of options. Fernald et al. (2017) identified 147 ECD assessments being used in LMICs and Munoz-Chereau et al. (2021) systematically review the psychometric properties of 34 assessments used in peer-reviewed publications about ECD in LMICs from 2010-2019. While some of these tools measure specific aspects of child development such as vocabulary size or communicative skills, many of the most popular assessments assess a broader construct covering multiple domains of child development. Domain and content coverage varies by the specific assessment in question, but Fernald et al. (2017) group domains of development into six broad groups: “cognitive skills,” “self-regulation, effortful control, and executive function,” “language skills,” “motor skills,” “social and emotional development,” and “pre- and early academic skills.”

3.1. Modalities of assessment

Two main types of assessments are used to generate data on ECD outcomes: direct assessment and adult report. Adult report assessments involve interviewing an adult that knows the child well (typically a parent or caregiver) and asking questions about what the child can and cannot do. In contrast, direct assessments engage children directly in a series of standardized activities that elicit behaviors indicative of their development. There are pros and cons of each modality and an ideal approach combines information from each (Renk, 2005).

Adult reported measures are typically implemented through an interview with the caregiver or a written survey (Rao et al., 2021). Adults can draw on their daily experiences with the child and reflect on what children can do in general. An adult therefore could report that a child is able to do something at home, even if they are not able to demonstrate the skill on the day of assessment. Adult reported measures might be particularly advantageous when attempting to measure ECD with very young children who are not yet able to engage in more complicated activities. At the same time, adult report measures can be vulnerable to response, recall, and social desirability bias (Bennetts et al., 2016; Law & Roy, 2008).

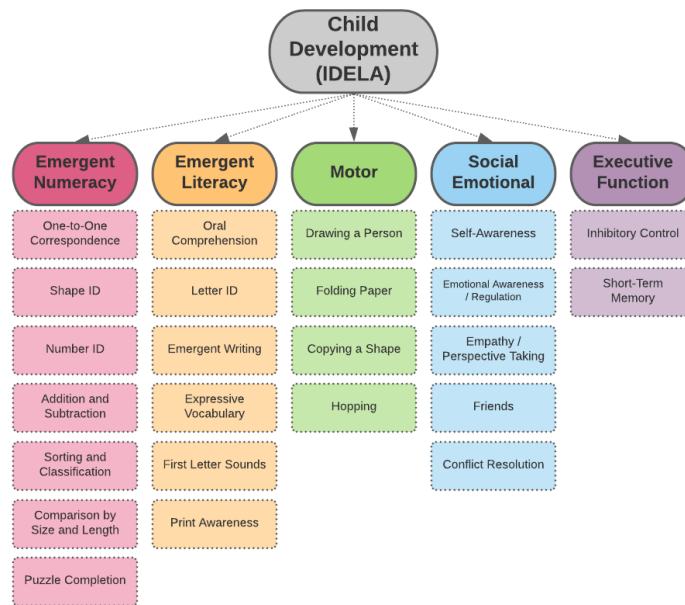
In principle, direct assessments provide a more standardized way to measure ECD through rigorous training of data collectors in order to achieve consistent administration procedures and understanding of scoring rules (Munoz-Chereau et al., 2021). Direct assessments are also able to design specific activities to elicit important skills (such as those in executive functioning) which adults may be less aware of and able to reliably report. At the same time, direct assessments place children in unfamiliar settings which could lead to construct-irrelevant reasons for low scores, and enumerators could be vulnerable to their own biases in scoring assessments (Bracken, 2007).

While each modality has its strengths and weaknesses, direct assessment is often considered the best approach to ECD measurement and has become a cornerstone of research and program and policy evaluation in LMICs (Fernald & Pitchik, 2019; Rao et al., 2021; Sachse & Suchodoletz, 2008). Despite its attractive attributes leading to an improved perception of its validity, direct assessments are expensive, time-consuming, and complicated to administer with reliability and fidelity (Gardner, 2000; Munoz-Chereau et al., 2021). For example, the Bayley Scales of Infant and Toddler Development, often considered a “gold standard” of ECD direct assessment, takes 50-90 minutes to administer (Albers & Grieve, 2007). While useful for detailed studies and evaluations, direct assessments are thus rarely used for broad-based population monitoring, which typically use adult-reported measures (Fernald et al., 2017; McCoy, Waldman, et al., 2018). For example, the ECDI2030, an adult report measure designed for integration into nationally-representative household multitopic surveys, takes fewer than 10 minutes to administer (Halpin et al., 2023).

3.2. The International Development and Early Learning Assessment (IDELA)

IDELA is one example of a holistic direct assessment of ECD that was developed in response to a paucity of tools that were feasible for use in LMICs (Pisani et al., 2015, 2018; Wolf et al., 2017). Save the Children began development of the IDELA in 2011 aiming to create a tool that was 1) relatively easy to implement, 2) focused on capturing the distribution of development rather than on screening for delay, and 3) tested skills that were broadly relevant across languages, cultures, and countries (Halpin et al., 2019; Pisani et al., 2015, 2015; Wolf et al., 2017). After considering 65 subtasks (activities) in field tests and in consultation with global subject matter experts, IDELA creators narrowed the tool to a selection of 24 subtasks for inclusion in the IDELA (Pisani et al., 2015, 2018). These 24 subtasks are organized into the “core” domains of Motor (4 subtasks), Social-Emotional (5 subtasks), Emergent Literacy (6 subtasks), Emergent Numeracy (7 subtasks) and two additional subtasks classified as Executive Function. Figure 1 identifies each IDELA subtask and the domain with which it is associated (Pisani et al., 2018).

Figure 1. *Domains and Subtasks of IDELA*



Note: Emergent Numeracy, Emergent Literacy, Motor, and Social-Emotional are consisted “core” domains of IDELA, with Executive Functioning a supplemental domain. Figure created by author based on Save the Children (2019)

Administration of the IDELA typically takes about 25-35 minutes and involves the use of materials such as a booklet of stimulus cards, locally appropriate materials for counting, and drawing materials and paper (Save the Children, 2019). Since IDELA was first tested by Save the Children, usage of the free and open-source tool has exploded, with implementation in over 75 countries by dozens of non-governmental organizations, inter-governmental organizations, academics and governments (Save the Children, 2024).

4. Creating short forms of assessments

As the name suggests, short forms of assessments comprise a subset of the items included in the full-length assessment. Short forms of direct assessments of ECD could be a means to derive some of the benefits of a direct assessment at a fraction of the cost and complexity. However, a reduced number of items in an assessment necessarily results in a decrease in reliability and precision of scores with increased measurement error (Frey, 2018). This decrease in reliability can still be acceptable, especially when short forms focus on group-level analyses and benefit from larger sample sizes than would have been possible with a longer assessment. The paragraphs below describe the “traditional” statistically motivated method of selecting items or subtasks for inclusion in a short form, illustrate why this approach is often insufficient and how users have addressed these limitations. Finally, it proposes an alternative framework which explicitly considers multiple dimensions of items or subtasks when creating a short form.

4.1. Statistically focused approaches to creating short forms

Historically, statistical properties of items were the primary consideration when creating short forms. Approaches differ in detail, but Classical Test Theory, Factor Analysis, and Item-Response Theory methods all identify item how strongly items intercorrelate (in the case of Classical Test Theory) or with a latent construct(s) of interest (for Factor Analysis and Item Response Theory). Researchers then select the items that are most intercorrelated or that best discriminate between individuals with higher or lower values of the underlying construct to include on short forms (Widaman et al., 2011). For a given number of items, this approach maximizes the reliability of the assessment and test information function. While

maximizing the statistical properties of short form, this approach also implicitly assumes that, other than their statistical properties, items are otherwise interchangeable. A focus on reliability can undermine the validity of assessments by, for example, only selecting items that cover a narrow range of a broader construct being assessed (Clark & Watson, 2019; John & Soto, 2007; Little et al., 1999). This “attenuation paradox” means that increasing the internal consistency of a short form can end up decreasing the validity of the scores generated (Boyle, 1991; Loevinger, 1954).

Smith et al. (2000) outline how naïve approaches have led to “sins” in short-form development. These range from failing to show that the short form preserves content coverage from the full form to failing to demonstrate that cost or time savings are worth the reduction in reliability (Smith et al., 2000). The assumption of interchangeability can be useful for creating short forms of written tests of a narrow construct but is untenable for measuring a broad multi-dimensional construct such as ECD. The subtasks on direct assessments such as IDELA attempt to capture development across a broad range of domains and there is a strong risk of artificially narrowing the construct of measurement by relying solely on statistical means. Similarly, the practical considerations of administering a direct assessment of ECD such as IDELA also vary widely. Some activities are easy for enumerators to learn, quick to administer, and require simple or few materials. Others require intensive training and feedback to achieve reliable administration, take more time to score, and require specialized materials. In short, relying solely on the statistical criteria of items can result in selecting a sub-optimal short form of a direct assessment of ECD that artificially narrows the object of measurement and that does not maximize improvements in its usability.

4.2. Incorporating non-statistical information and Automated Test Assembly

Aware of the limitations of a solely statistically focused approach and in an attempt to avoid creating a short-form with decreased validity or practicality, many developers of short form assessments incorporate information beyond the statistical properties of items and subtasks (Smith et al., 2000). For example, in selecting the child assessments for inclusion in the 1986 National Longitudinal Study of Youth, Baker & Mott sought to find assessments that were not only “highly reliable and valid,” but also

“inexpensive to administer,” “require[d] very little equipment,” and “recognized by the social science community” (Baker & Mott, 1989, pp. 47–48). Similarly, in developing short forms of patient-reported outcome measures, Cella et al. (2019, p. 539) “incorporated item statistics and input on the content from clinical experts” to select individual items that would have adequate content coverage and Salsman et al. (2020, p. 5) “used group consensus” with a panel of experts to examine statistical criteria and determine the final inclusion of items of a short form measuring “meaning and purpose in life.” These tool developers avoided selecting a sub-optimal set of items by incorporating qualitative information beyond the statistical properties of items to ensure validity and practicality of the instrument.

Automated Test Assembly is an approach that would allow for the systematic incorporation of non-statistical information in short form development and selection of an ideal short form. Automated Test Assembly is the state-of-the-art method to create tests that selects mathematically optimal items from an item bank against an objective function. The theory underpins the advantages that modern Computer Adaptive Testing brings to shorten tests, assists test developers in creating equivalent parallel forms, and can be used to select optimal items for a short form (Linden, 2005). While Automated Test Assembly was not specifically designed to handle non-statistical criteria such as the conceptual strength of an item, administration length, or administration complexity, any requirement that can be coded quantitatively and made into a constraint can be used to derive optimal tests.

For advanced test developers seeking to create a short form across multiple defined optimization criteria, Automated Test Assembly is the ideal approach. For example, given constraints that a test must be no longer than 10 minutes, should not require the use of any specialized materials, must contain at least one item from each domain, and should maximize information for children 2 SD below the mean in terms of development, Automated Test Assembly would be able to select the best items or subtasks to achieve this purpose. However, the strengths of Automated Test Assembly also make it complex for many users to operationalize. The mathematical programming required to implement Automated Test Assembly solutions is complex, and while simpler-to-use packages exist for more common use-cases of Automated Test Assembly, customizing these to incorporate criteria beyond the statistical properties of items is

complicated (Becker et al., 2021). Even when programming is not a barrier to implementation, translating qualitative desires about the attributes of a short form into strict mathematical constraints is challenging. A user may want to minimize the length *and* the use of specialized materials, but would accept various tradeoffs between the two (e.g., an eight-minute test using one set of specialized materials might be equally acceptable to a ten-minute test without any specialized materials). Formalizing these tradeoffs into mathematical expressions that can be used by Automated Test Assembly is a non-trivial task.

4.3. A multi-dimensional human-centered framework

Rather than focusing exclusively on the statistical properties of items and without the complexity that Automated Test Assembly brings, this paper proposes a framework for creating short forms of direct assessments of ECD that considers three dimensions as described in Table 1. The *statistical* dimension is analogous to the information used in traditional methods of creating short forms and prioritizes items with the highest reliability to minimize measurement error. The *practical* dimension considers the ease of administration and scoring of each item along with the difficulty of training requirements and prioritizes the use of quick and easy-to-score items. The *conceptual* dimension seeks to ensure included items are representative of the same construct measured by the full tool and that captures the most important skills within the construct. Optimal short forms of direct assessments of early childhood maximize the use of the most reliable and highest information items (statistical dimension), minimize the use of complex and time-consuming items (practical dimension), and preserve the validity of the assessment by ensuring that selected items are representative of the constructs intended to be measured by the full assessment (conceptual dimension).

As shown in Table 1, each dimension requires the use of a different source of data to evaluate items for inclusion. The statistical dimension requires item-level data from a relevant population to judge the information generated. The practical dimension requires data on the time required to administer each item along and benefits from the judgements about the complexity of training and scoring. The conceptual dimension requires the systematic collection of the opinions of tool users and Subject Matter Experts

(SMEs) to identify the most important items to retain on a short form and help create decision rules for what the short form must include.

Table 1. Dimensions for consideration when selecting items for a short form

Dimension	Goal	Data required
Statistical	Maximize precision and reliability of scores, minimize standard error of measurement	Item-level data from full assessment
Practical	Minimize item administration time, complexity of training and scoring, and use of special materials	Data on length and complexity (e.g., required materials) of administration, opinions of scoring and training complexity from tool trainers
Conceptual	Ensure items selected are representative of construct measured by full assessment	Opinions of subject matter experts and tool users

Figure 2. Process of selecting items/subtasks for inclusion on a short form using a multi-dimensional framework



Figure 2 describes the steps necessary to use this framework in practice. The creator of a short form must first **collect** relevant data to assess the tool according to the three dimensions and **consolidate** data into actionable decision rules that guide the selection of the items or subtasks for the short form. Next, users should **visualize** the relevant information across the three dimensions both individually and jointly. Finally, users must **decide** on the subset of subtasks or items to include based on this information.

Importantly, while data-driven, the use of this framework should not be considered a purely empirical process that is fully replicable. The primary goal of this process is to synthesize information across important aspects of an assessment and provide a structured manner to analyze and visualize the tradeoffs of including various items or subtasks. At the final step, the decision to select a particular set of items or subtasks is a qualitative choice and rests on human judgment.

This approach contrasts with Automated Test Assembly, which can give an optimal solution, but requires the hard-coding of constraints by the test creator. In many ways, the proposed framework attempts to systematize the process that many short form creators have used implicitly by providing a clear structure through which to evaluate items and incorporate expert opinion into the decisions made to include specific items (Baker & Mott, 1989; Cella et al., 2019; Salsman et al., 2020).

5. The current study

Relative to many ECD assessments, IDELA is quick to administer and easy to train on. However, with an average assessment length of 25-35 minutes per child and various materials required for different subtasks, it still represents a considerable use of time for each child and requires a four-to-five-day training to ensure reliable administration. As such, IDELA is infeasible to use for regular monitoring or large-scale data collections. Given its wide acceptance and extensive use globally and its nature as a direct assessment with significant variation in the complexity of administration, it represents an ideal candidate for creating a short form that would support additional use cases.

The current study attempts to apply the multi-dimensional human-centered framework introduced above to develop a general-purpose “balanced” short form of the International Development and Early Learning Assessment (IDELA) that considers information from each of the dimensions. The current study is driven by two primary research questions:

- 1) How do the subtasks of IDELA vary in terms of their statistical, practical, and conceptual dimensions?
 - a. What subtasks would a “balanced” short form IDELA include?
 - b. What subtasks would a “traditional” statistically-focused short form IDELA include?

- 2) How do results generated from a “balanced” IDELA short form compare to the full assessment and a “traditional” short form?

The goal of the multi-dimensional framework for creating short forms described above is to provide the test creator with more complete information with which to make a decision on which items to include in the short form. As such, the “results” of RQ1a also include using the information generated by this process to create a series of decision rules and applying them to IDELA to select subtasks using human judgement.

6. Method

To answer RQ1, I apply the multi-dimensional framework using data illustrated in Table 1 to collect and analyze data. I rely on a large item-level dataset of observations from a diverse set of countries to first determine whether IDELA data can reasonably be modeled as representing a unidimensional construct and then to estimate the information and reliability of subtasks using a Graded Response Model (Samejima, 1968). For the practical dimension, it estimates how long each subtask takes to administer by coding 13 videos of English-language IDELA administrations that were created in the United States for training purposes. Finally, for the conceptual dimension, survey data from 14 SMEs are collected to understand the degree to which IDELA items are representative of the constructs they are designed to measure and the desired properties of a short form IDELA.

Having collected data on the properties of RQ1, I attempt to answer RQ1a by combining information across dimensions and selecting subtasks for a balanced short form in consultation with the IDELA creator. I also select subtasks for a “traditional” short form (RQ1b) at the same time. The psychometric properties of these two short forms are then compared with the full IDELA (RQ2).

6.1. Measures

6.1.1. Measures of statistical properties

The statistical properties of IDELA are assessed through a large dataset of child-level data on IDELA. IDELA contains 24 subtasks. While item-level data is available for the assessment, this analysis focuses on subtask scores as the source of data on the statistical dimensions to improve theoretical

comparability across datasets and also for the practical consideration that keeping individual items of a subtask does not reduce administration burden nearly as much as removing entire subtasks.

Most IDELA items are nearly identical across administrations and only differ in the language of administration and use of locally relevant materials (e.g., counting with a set of contextually familiar small objects). This is not true with some of the language subtasks. In the “Letter Identification” subtask, for example, the child is asked to identify a series of letters or, depending on the alphabet, characters or syllabograms. These 20 letters are drawn from the most common in the language of administration and are presented not in alphabetical order. Instead, letters are split into the 1-10th most common and ordered randomly, and then the 11th-20th most common and ordered randomly. Thus, `letterid1`, the variable representing whether the child identified the first in the subtask letter correctly, is not conceptually distinct from `letterid2` or `letterid3` or others. The construct that the entirety of these variables represents is similar across languages, but the item-level construct representation is inconsistent. Any findings about the relative information given by a particular letter variable would likely be idiosyncratic and not represent true differences in item information. Similar conceptual issues arise when considering the performance of the “First Letter Sounds” sounds subtask, where children are asked to identify words that start with the same phoneme, a task that very necessarily varies by language both in the words used, and the phonemes assessed. Analyzing IDELA data at the subtask level also helps to balance the items in assessment by domain by lending equal weight to subtasks with more and fewer numbers of items.

It also makes sense for this analysis to focus on subtasks for practical reasons. The materials, administration approach, and training for a subtask apply to all the items in the subtask. Individual items within a subtask may well differ in their reliability, but once a subtask is included, along with the training and materials that it requires, the incremental cost of administering an additional item is minimal. The goals of creating a Short Form IDELA are thus more easily realized by preserving and dropping entire subtasks than preserving or dropping individual items.

For these theoretical and practical reasons, this analysis treats IDELA subtasks as the unit of analysis. To do so, I calculate “percentage correct” for each subtask by dividing the total score of the

items in the subtask by the total possible score. This percent correct scoring leads to all subtasks scored on a scale of 0-100%, but the number of increments within subtasks depends on the total possible score. For “Conflict Resolution,” which has just two items scored correct/incorrect, the possible scores are 0, 50, and 100. For subtasks such as “Number Identification,” which has 20 items, the possible scores are any number 0-100 divisible by 5.

6.1.2. Measures of conceptual and practical properties

To assess IDELA’s conceptual and practical properties, I use two measures: videorecorded administrations of IDELA that were taken for training purposes and a survey provided to users of IDELA. The video recordings were used to extract the starting and stopping time of each subtask and calculate the length of time spent administering each subtask. These start/stop times included any transitions between subtasks to help include the effects of complexity introduced due to using special materials.

A survey was conducted among IDELA SMEs to gather information on both what was important for the short form to contain conceptually as well as practical guidelines for its use. The full survey instrument is provided in Supplemental Materials Annex A: IDELA User Survey.

This survey had two main sections. The first section gauged the SME’s enthusiasm for a Short Form and the desired parameters regarding its length, domain coverage, and the importance of its psychometric properties vs. other factors. After being asked about the general attributes of the short form, SMEs were also asked to identify the subtasks of IDELA that they believe most represented the domains included in the assessment. For each of the core domains, users selected two subtasks they felt were most representative of that domain, and one subtask they felt was least representative.

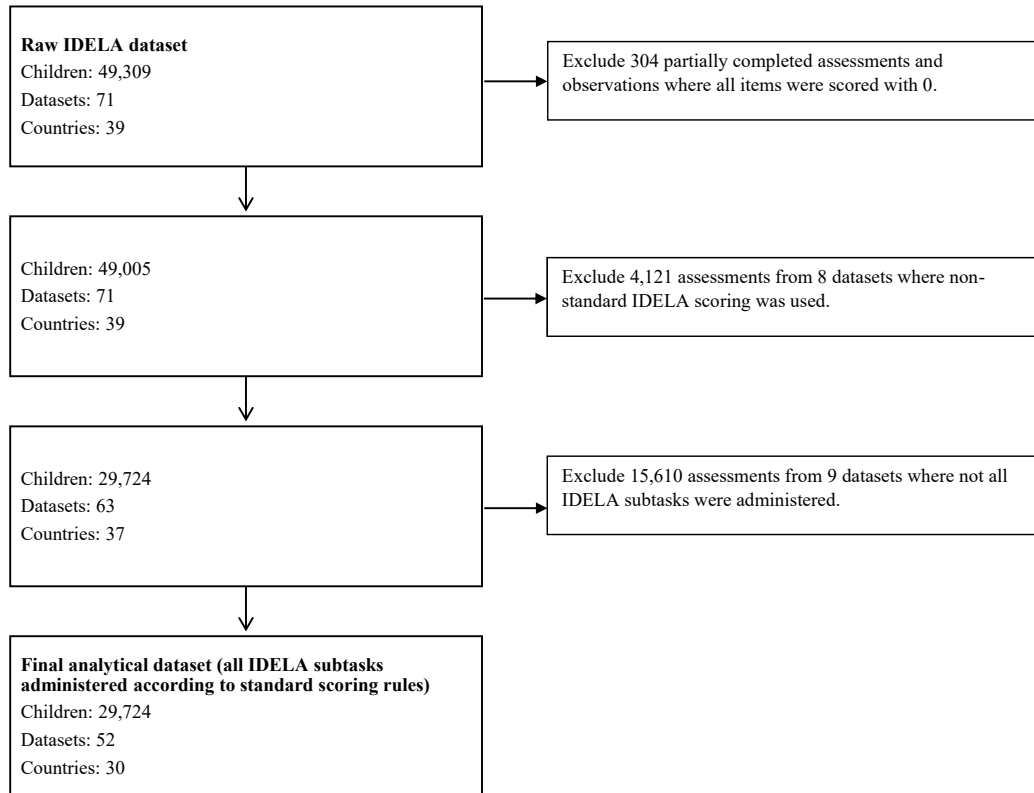
6.2. Participants

6.2.1. International IDELA dataset used to assess statistical dimension

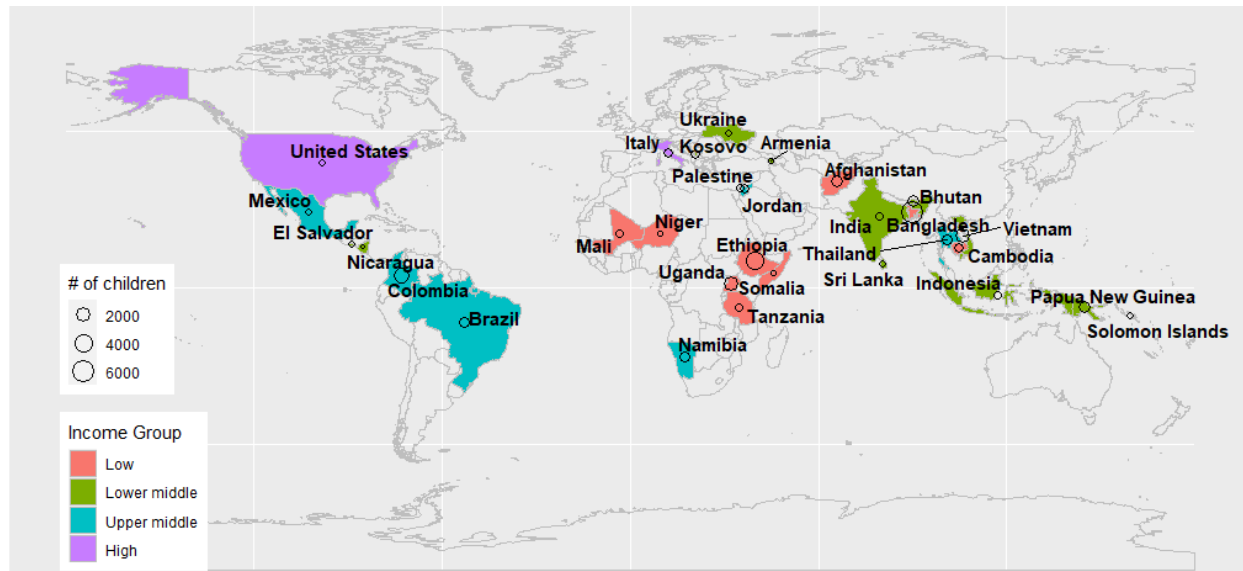
Save the Children provided a large item-level dataset which includes 49,309 anonymized observations from 71 separate data collections conducted by over a dozen organizations in 39 countries between 2015 and 2020. As noted above, adaptation and contextualization of IDELA is often required prior to administration in a new context. This helps ensure that contextualized versions of IDELA

measure generate contextually relevant data. However, when items or subtasks are extensively contextualized, the subtask may measure a substantively different construct. To limit the effect of these contextualizations, this analysis only includes the observations of IDELA which are most theoretically comparable, meaning that they were administered and scored in nearly the same manner.

The exclusion rules and how many observations are excluded are detailed in **Figure 3**. Out of the original 71 datasets, 17 were excluded from this analysis because 1) they did not administer all 24 IDELA subtasks or 2) used non-standard scoring patterns for some subtasks that make equating scores impossible. For example, two datasets from Ghana were excluded because they did not include the “Hopping” subtask, a dataset from Rwanda was excluded because it omitted several items from the “Personal Awareness” subtask, and three datasets from the Middle East were excluded because they used a reduced number of letters and numbers in the “Letter ID” and “Number ID” subtasks. In addition to discarding these data, individual incomplete administrations within datasets and observations where all items were scored as zero were dropped. This ensures that all data included in the analysis has a comparable set of subtasks that were all scored in a similar way.

Figure 3. Process of sub-setting IDELA dataset to analytical sample

These strict rules led to dropping about 42% of the original dataset, but left a large sample of 29,724 observations from a diverse set of 30 countries as shown in **Figure 4**. The biggest differences between the raw dataset are regarding the regional representation. There are relatively fewer observations from Sub-Saharan Africa (27.6% instead of the original 44.0%) and Middle East and North Africa (3.7% instead of the original 6.8%) and a relative increase in representation from East Asia and Pacific (16.8% instead of the original 11.8%), South Asia (31.7% instead of the original 22.3%), and Latin America and Caribbean (14.4% instead of 10.4%). Most of the countries included in the analytical sample are in the World Bank's Low (9) and Middle Lower (13) income groups, but it also includes six Upper Middle and two High Income countries.

Figure 4. IDELA data sources for statistical dimension by country and income level ($n=29,724$)

6.2.2. Participants in data used to assess conceptual and practical dimensions

A total of 13 videorecorded administrations of IDELA were used to estimate the length and complexity of the IDELA subtasks for the practical domain. All videos were of English-language administrations and recorded in the United States for training purposes.

For the user survey, the listserv of the IDELA Network website was used to contact the registered users and provide them with the opportunity to anonymously respond and voice their priorities for a Short Form IDELA. A total of 14 IDELA users responded to the SME survey. The survey did not collect demographic information on users, but users of the IDELA Network include ECD practitioners designing and implementing interventions, academics, and applied researchers who have experience administering and training IDELA itself or using IDELA data in their research. Out of 14 participants, 9 reported having personally led an IDELA training themselves and all but one reported having personally been involved in the collection of IDELA data previously.

6.3. Analysis

The analysis begins by answering RQ1 and identifying the statistical, practical, and conceptual properties of each IDELA subtask. To identify subtasks for inclusion in a “balanced” IDELA short form I combined information across these dimensions (RQ1a) and used only the statistical dimension to create a

comparative “traditional” short form (RQ1b). After creating each of these short forms, I compare their psychometric properties (RQ2) against the full IDELA assessment.

6.3.1. Statistical dimension

The analysis for the statistical dimension relies on Item-Response Theory, with a Graded Response Model as the primary model used to assess subtask performance (Samejima, 1968). Before applying this model, I begin by exploring the Classical Test Theory properties by examining the correlations between subtasks as well as the overall Cronbach’s alpha of the assessment. An assumption of all IRT methods is that of unidimensionality—that all items on an assessment are measuring the same construct (Lord, 1980). From a theoretical standpoint, ECD is certainly not unidimensional and is generally agreed to be a complex multi-dimensional process. Indeed, the IDELA was specifically created to capture development across multiple domains of development. Despite this seeming violation of a core assumption of Item Response Theory, previous studies of IDELA in several countries noted that while a bifactor model consisting of multiple domains is a superior fit to the data, that a “[u]nidimensional model also fit the data from each country quite well”, and that a future avenue that might facilitate cross-country comparison would be to simplify the conceptual model of the assessment to a single factor (Halpin, et al., 2019, p.13). Further, short forms of hierarchical constructs are often constrained by their nature to produce only a single overall score given a lack of discriminant validity on a shorter assessment (Clark & Watson, 2019; Smith et al., 2000). This is generally accepted for the purposes of monitoring and evaluation other ECD assessments with the ECDI2030 and the GSED relying on a simplified unidimensional score (Cavallera et al., 2023; Halpin et al., 2024; van Buuren & Eekhout, 2024). Based on this, a principal factor analysis is first conducted to test whether a unidimensional simplification might be reasonable.

After establishing that the data are adequately unidimensional, the analysis continues by fitting a Graded Response Model and estimating parameters for each subtask as shown in equation 1. This equation allows us to estimate the conditional probability that the observed response X for child i on

subtask j is greater than or equal to score level k given θ , the underlying unidimensional construct assessed by the items on the assessment, which in this case is early learning and development.

$$P(X_{ij} \geq k | \theta_i) = \frac{e^{a_j(\theta_i - b_{jk})}}{1 + e^{a_j(\theta_i - b_{jk})}} \quad (1)$$

For each subtask, a shared a (discrimination) parameter and b (difficulty or location) parameters for each score-level are estimated using Stata 17.0 (StataCorp, 2019). The a parameter describes the magnitude of change in the log-odds of scoring at or above any given score level in the subtask associated with a one-unit change in a child's θ . Score levels are the “percent correct” scored on the subtask, as such, the b parameter for each score-level of a given subtask is the θ at which a child has a 50-50 chance of scoring at or above that percentage correct on the subtask.

This analysis focuses on the a discrimination parameters rather than the b parameters as the primary statistical criteria for two primary reasons. First, since the objective is to create a test that maximizes information across the ability spectrum, the a parameters show which subtasks provide the largest amount of information, regardless of where that information is produced. Second, the a parameter is a single summary of information whereas each subtask has multiple b parameters. As such, the a parameters provide a better single summary of information. For other goals, it would be more appropriate to focus on the b parameters—this is addressed in the Discussion.

6.3.2. Practical and conceptual dimensions

The results of the user survey are used in two ways. Summaries of responses from the first section are used to create decision rules about the short form of the assessment, such as the goal length of administration of a short form IDELA and the importance of full domain coverage in subtasks in the short form. The second half of the responses are used to summarize information about the relevance of individual subtasks. An “Subject Matter Expert” score is calculated for each subtask. The Subject Matter Expert score is simply the number of respondents who indicated the subtask was the most important and subtracting the number of respondents who thought that the subtask was the least important within the domain.

6.3.3. Combining information across dimensions and selecting subtasks for short form

As described in Table 1, the multi-dimensional human-centered framework for creating a short form introduced in this paper provides a structured process for collecting relevant information and assessing the strengths and weakness of items or subtasks of an assessment across three dimensions. However, and in contrast to other approaches to short forming such as Automated Test Assembly, this approach does not provide a singular solution for the “best” short form given these dimensions. Instead, it allows the user to create a series of visualizations combining information across dimensions to illustrate the tradeoffs of including different subtasks. For this paper, **R** and **ggplot2** was used to create visualizations of information across dimensions (R Core Team, 2021; Wickham et al., 2019). The ultimate decision on the number of items or subtasks to include and inclusion or exclusion of individual subtasks must be made with human judgement.

For this paper, the human element was a joint meeting between the author and Dr. Lauren Pisani, where we comprehensively reviewed the evidence presented in the results section and discussed the tradeoffs of including each of the subtasks and created a balanced general-purpose IDELA Short Form (IDELA-SF). For comparative purposes, a “traditional” short form was also created by ignoring all information except for the *a* parameters from the statistical dimension and selecting the eight subtasks with the highest discrimination parameters.

6.3.4. Comparison of short forms

After selecting the subtasks for the balanced and traditional short forms, the psychometric properties are compared along with the parameters for the full assessment using secondary Graded Response Models for each selection of subtasks. The test information function and conditional standard errors of measurements generated by the full IDELA, traditional short form, and balanced short form are used to assess the loss of precision on short forms using different selection criteria and compare these results to differences in the conceptual and practical domains. To supplement the psychometric analysis of both short forms, the domain coverage, average SME rating of subtasks, and estimated administration length for each form is also estimated.

7. Results

To answer RQ1, information from each of the dimensions is presented individually below. Then I combine information across domains and select subtasks for a “balanced” short form (RQ1a) and a comparative “traditional” short form (RQ1b) before comparing the results of each to the full IDELA assessment.

7.1. Statistical dimension

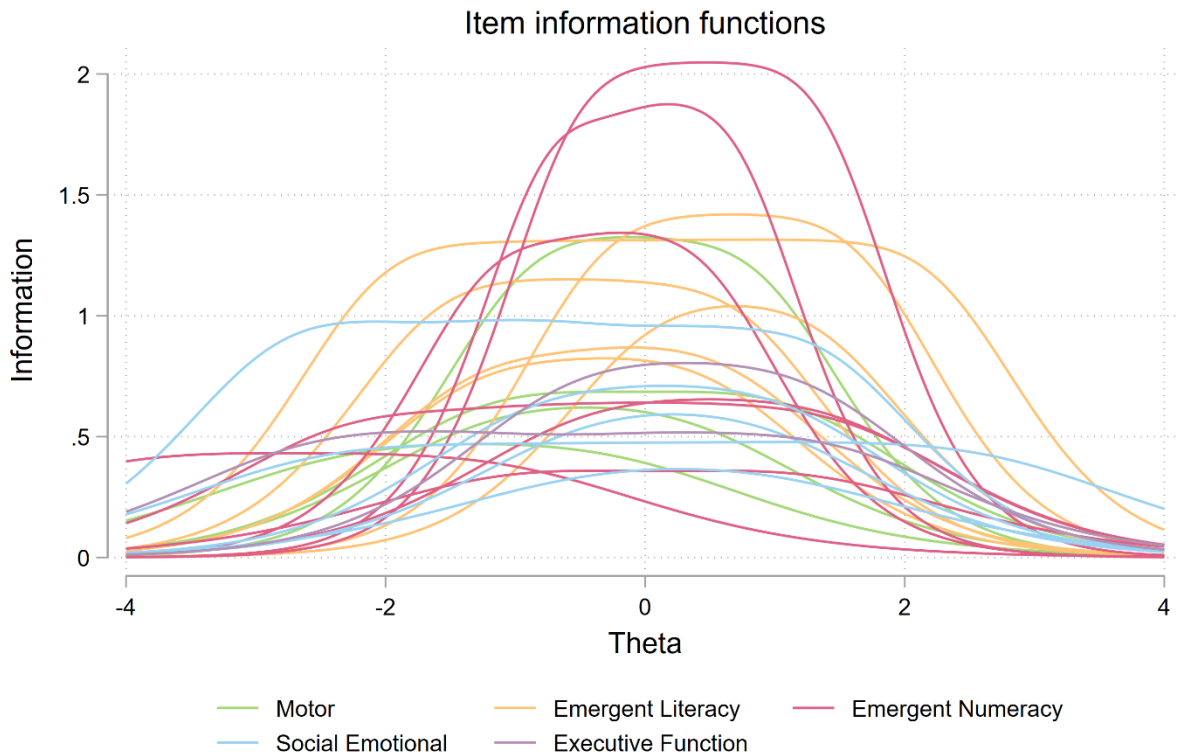
Before fitting the GRM to the data, I first explore the Classical Test Theory statistics and dimensionality of scores by examining the correlation between subtasks, the Cronbach’s alpha of the IDELA assessment, and the results of a principal components analysis. Full results are presented in Annexes A and B. The correlation matrix of all subtasks revealed that all subtasks are significantly correlated with each other ($p < .001$). The range of correlations ranged substantially from a modest correlation of $\rho = .226$ (between the “Conflict Resolution” and “Puzzle Completion” subtasks) to a strong correlation of $\rho = .758$ (between Letter Identification and Number Identification). These strong correlations are an encouraging result that suggest that all IDELA subtasks are related to a common underlying factor. This finding is further reinforced examining the Cronbach’s alpha reliability coefficient for the assessment. The 24 subtasks had a Cronbach’s alpha of .937, indicating a high degree of reliability and indicating that approximately 94% of variance in IDELA scores is attributable to true score variance in a single ECD construct. Importantly, variance in estimates of development includes the association with age, which is strongly correlated with each item. Full test results including item-test and item-rest correlations are presented in **Supplemental Materials Annex B: IDELA Subtask Descriptives and Correlations**.

To empirically test whether IDELA subtasks could be adequately represented as a unidimensional construct, I conducted both a Principal Components Analysis and a Factor Analysis and show results in Annex B. Visual inspection of the scree plot shows an unmistakable “elbow” after the first factor, strongly suggesting that a single factor effectively explains an enormous proportion of variance in scores (Wilcox, 2017). A single component can explain over 42% of variance across 24 distinct subtasks,

whereas the second component explains just 5.4% of residual variance. I conclude that modeling IDELA scores as a single factor is a reasonable approach.

Next, I fit a Graded Response Model to all 24 IDELA subtasks. **Figure 5** plots the individual Item Information Functions for the full IDELA assessment and full parameter estimates for each subtask are provided in **Supplemental Materials Annex D: IDELA Graded Response Model Parameter Estimates**. Overall, all subtasks appear to have reasonably high discrimination a parameters—indicating that, as a child’s log odds of scoring at or above any of the subtasks score points is strongly related to θ . However, as shown in **Figure 5**, there are some large differences between the highest and lowest information subtasks. Overall, subtasks from the Emergent Literacy and Emergent Numeracy domains had the highest information peak parameters. This is not a surprising result given that we found that subtasks from these domains correlated more strongly with each other than with subtasks from the Social Emotional, Motor, or Executive Function domains.

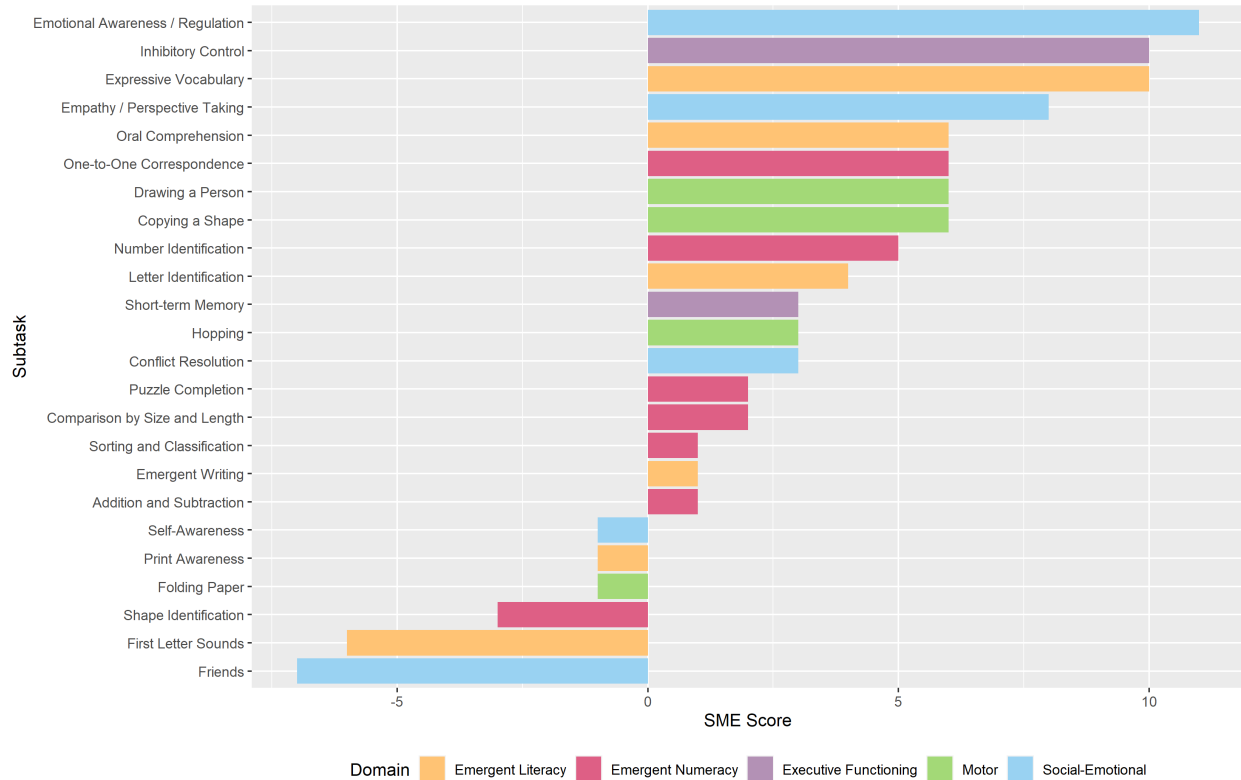
Figure 5. Statistical dimension: Subtask Item Information Functions by domain



7.2. Conceptual dimension

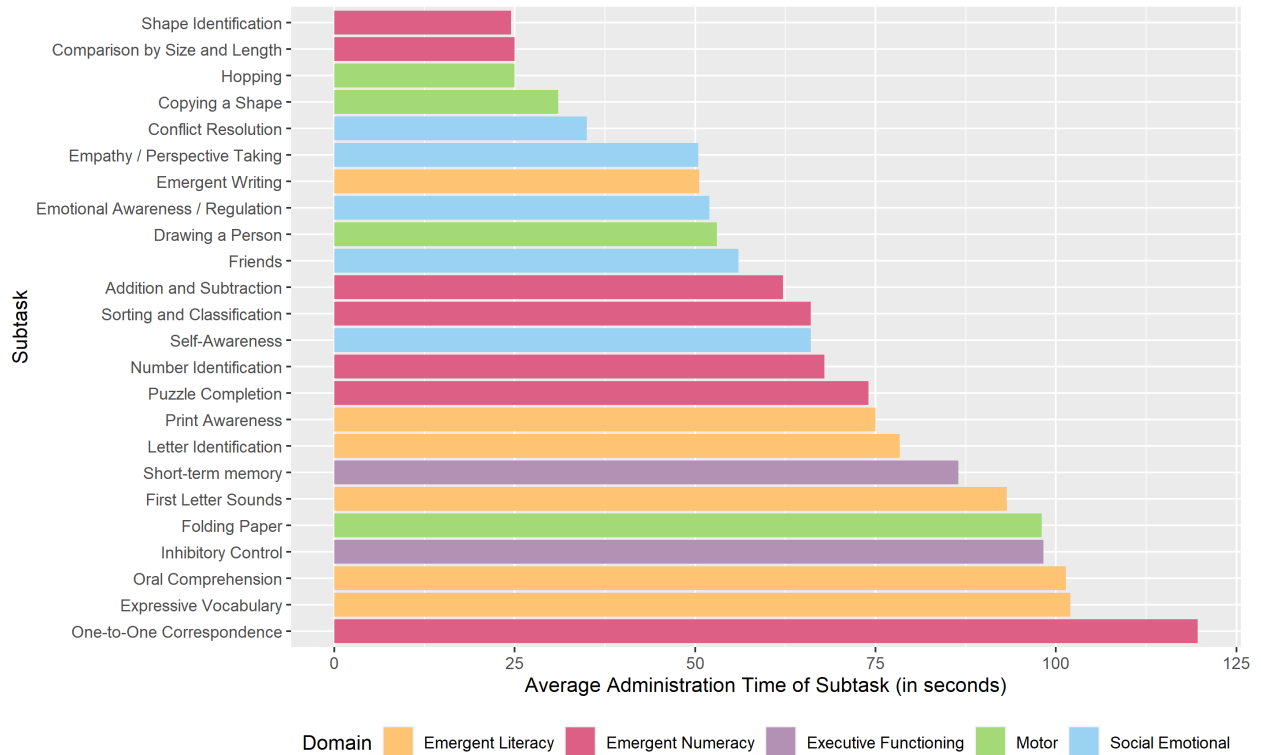
According to the survey, there was strong agreement that a short form IDELA should contain subtasks across each domain of the assessment, even if it meant sacrificing a degree of reliability. Out of 14 SMEs surveyed, all but one agreed or strongly agreed that “It’s very important that a Short Form IDELA includes subtasks from ALL domains of the assessment.” There was less agreement about how long the short form should take, with seven SMEs agreeing that “It’s very important that the Short Form takes less than 10 minutes to complete,” five disagreeing, and two expressing no opinion.

The second part of the survey focused on the conceptual importance of each subtask to the domain to which it belongs. The results of this analysis are presented in **Figure 6**. As the figure shows, there were considerable differences in how SMEs rate subtasks. For example, within the Social Emotional domain, the “Emotional Awareness/Regulation” subtask was included by 11 of 14 SMEs as one of the most important subtasks (and none indicating it was the least important), whereas 8 SMEs considered the “Friends” subtasks to be the least important in the domain (with just one SME indicating they thought it was one of the most important subtasks). There was also strong agreement that the “Expressive Vocabulary” subtask was most important to the Emergent Literacy domain and the “First Letter Sounds” least important. The Motor and Emergent Numeracy domains had less striking patterns of preference among SMEs, and the magnitude of differences within these domains was smaller.

Figure 6. Conceptual dimension: Subject Matter Expert scores of IDELA subtasks

7.3. Practical dimension

The primary metric used to quantify the practical dimension is the average length of administration. As shown in **Figure 7**, subtask administration length varies considerably across subtasks. Taking less than half a minute on average are the “Shape Identification,” “Comparison by Size and Length,” and “Hopping” subtasks, making them ideal candidates for inclusion considering the practical dimension. On the other hand, the “One-to-One Correspondence” subtask took nearly two minutes on average, about four times as long as the quickest subtasks.

Figure 7. Practical dimension: Average length of administration of IDELA subtasks

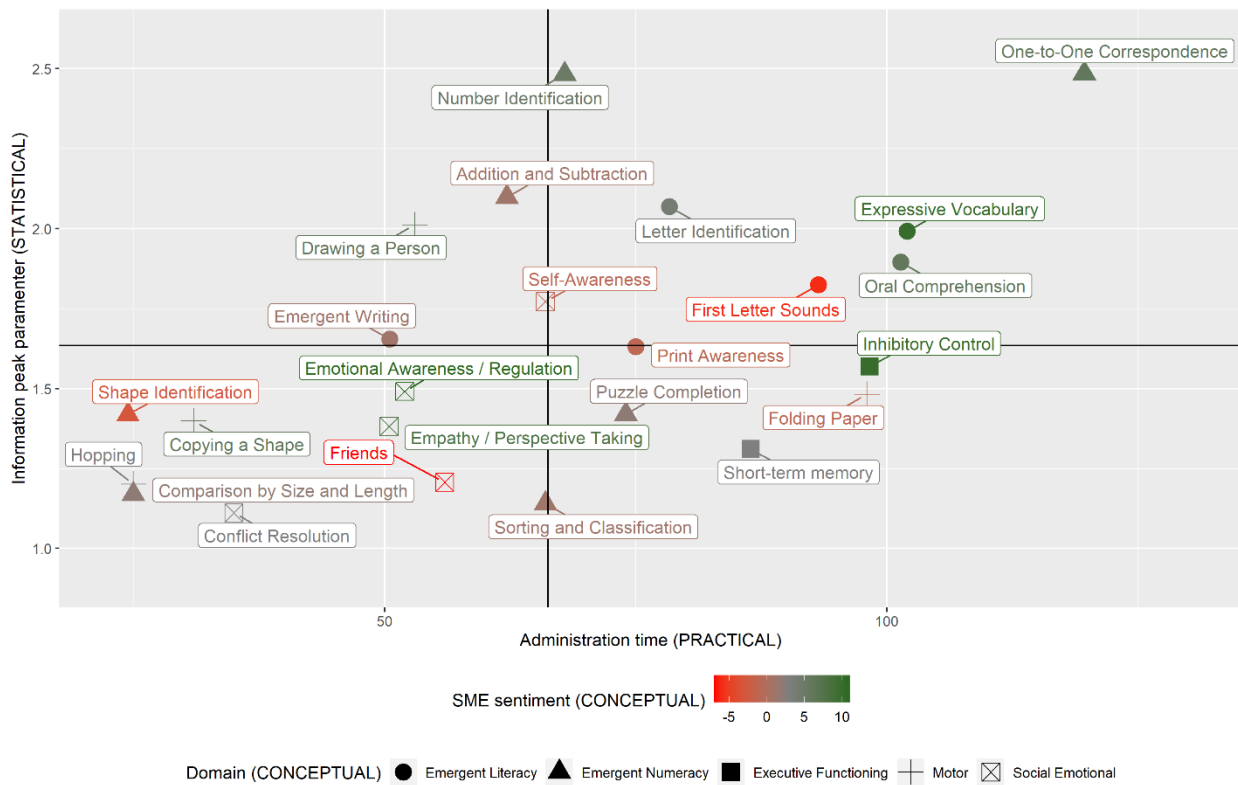
The length of administration of subtasks is an easy-to-quantify metric to gauge the practical dimension but does not capture the full range of practical considerations. Other considerations in the practical dimension are the special materials required to administer subtasks such as stimulus cards or objects to count, the complexity of administration rules and the time required to train data collectors, and the difficulty of accurately scoring responses. I did not attempt to quantify these other aspects of complexity, but discussions with the tool developer and IDELA master trainers identified the “Inhibitory Control” and “First Letter Sounds” subtasks as particularly difficult to train and reliably score due to complex multi-stage administration for the “Inhibitory Control” subtask and a lack of familiarity of phonemic awareness (as opposed to letter recognition) for many enumerators for the “First Letter Sounds” subtask.

7.4. Combining information across dimensions

The information across these three dimensions is combined into a single chart in **Figure 8**. This figure visualizes the tradeoffs for including each subtask and aids decision making when selecting

subtasks for inclusion in a balanced short form. In this figure, the y-axis represents the statistical dimension by plotting the a parameters of each subtask. Subtasks with higher a parameters capture more information about the developmental status of the child (the mean value of a parameters is shown with the horizontal black line). The x-axis attempts to capture the practical dimension by plotting the average time of administration of each subtask. Subtasks with lower administration time are easier and less complex to administer and score (the vertical black line represents the average time of administration). The conceptual domain is represented with the shape and color of points. The color of each subtask represents the SME sentiment score of the subtask—greener subtasks were deemed more important to include in a short form whereas redder subtasks were rated as least important to include. Finally, the domain of each subtask is also represented by the marker shape, helping ensure that each domain is represented in a balanced short form.

Figure 8. Combination of statistical, practical, and conceptual Dimensions of IDELA subtasks



7.5. Selecting subtasks for a balanced short form

Unlike Automated Test Assembly, the multi-dimensional human-centered framework introduced in this paper does not automatically select the best items for a short form. Instead, it helps organize information and quantify the tradeoffs of inclusion of subtasks or items in a short form. The ultimate inclusion of specific items or subtasks relies on a human decision. To answer RQ1a and create a “balanced” short form of the IDELA, the author worked with the IDELA creator to contribute this human element. Based on a review of the information generated, a set of decision rules was created to guide the selection of subtasks. The balanced IDELA Short Form should:

- 1) include an equal number of subtasks from each of IDELA’s core domains,
- 2) take no longer than 10 minutes to administer,
- 3) preference subtasks deemed conceptually more important,
- 4) and preference subtasks with higher discrimination parameters.

The decision to include some subtasks in the balanced short form was straightforward as they had positive attributes across each of the three dimensions. For example, “Number Identification” was deemed highly important by SMEs, had extremely good discrimination parameters, and was average in terms of administration time. However, other cases required assessing tradeoffs. For example, we decided to select an equal number of subtasks from each of the core domains of IDELA, but elected not to include any subtasks from the supplementary Executive Function domain. This was due to the practical demands in terms of the time of administration (“Inhibitory Control”) and sub-optimal statistical and conceptual properties (“Short-Term Memory”). In another judgement, we decided to include the “Empathy/Perspective Taking” subtask over the “Self-Awareness” subtask for the Social Emotional domain. Despite the superior statistical properties of the “Self-Awareness” subtask, the conceptual importance of “Empathy/Perspective Taking” was rated much more important by SMEs. Finally, despite the fact that it took more time than any other subtask in the analysis, the “One-to-One Correspondence” was selected for inclusion given its superb SME score and statistical properties.

After reviewing the above evidence and considering the tradeoffs of including each of the subtasks, a balanced general-purpose IDELA Short Form (IDELA-SF) is proposed. IDELA-SF includes

the “One-to-One Correspondence” and “Number Identification” subtasks from the Early Numeracy domain, the “Letter Identification” and “Expressive Vocabulary” subtasks from the Early Literacy domain, the “Drawing a Person” and “Hopping” subtasks from the Motor domain, and the “Emotional Awareness/Regulation” and “Empathy/Perspective Taking” subtasks from the Social-Emotional Domain. The balanced IDELA-SF has equal representation from the core domains of IDELA, should take approximately eight minutes to administer, and selects items that were statistically valuable, easy to administer, and conceptually important.

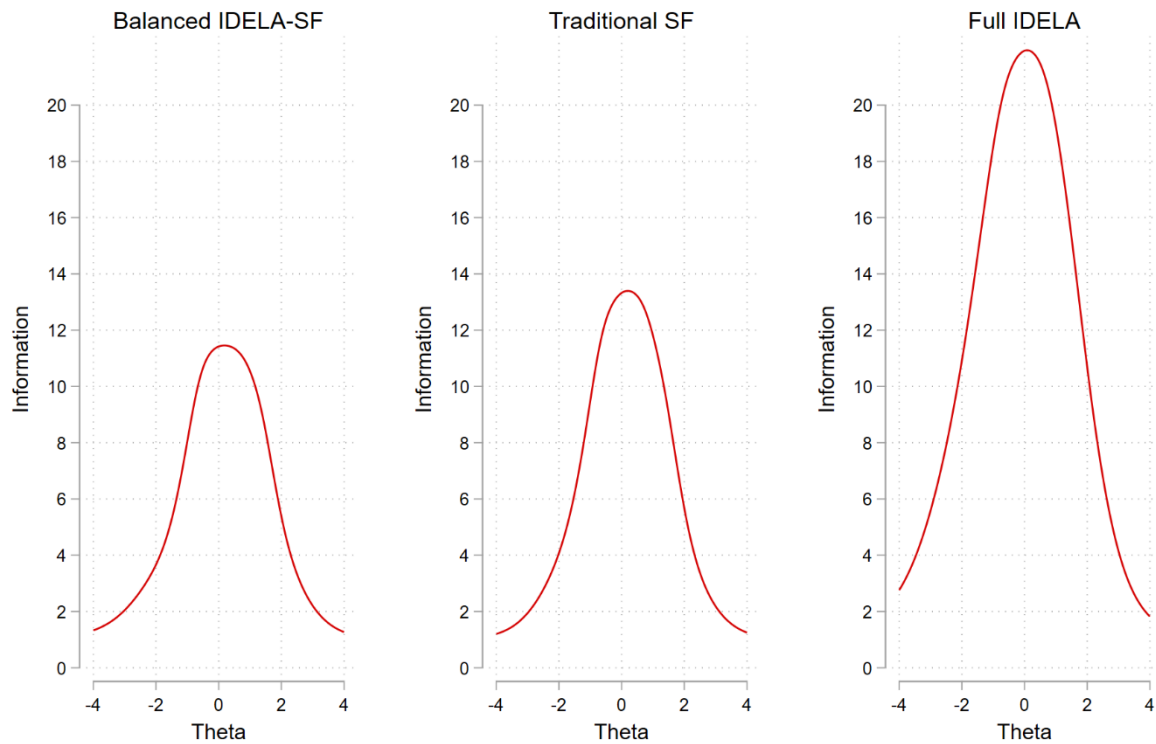
To assess the difference that using this framework made in terms of choosing subtasks for inclusion, a “traditional” short form considering only the statistical dimension of subtasks was also generated. **Figure 9** illustrates the differences in composition between the two short forms. Selecting the eight subtasks with the highest discrimination parameters would result in changing three of the eight subtasks compared to the balanced short form proposed above. This traditional short form includes the “First Letter Sounds” and “Oral Comprehension” subtasks from the Emergent Literacy domain and the “Addition and Subtraction” subtask from the Emergent Numeracy domain while discarding the “Hopping” subtask from the Motor domain and the “Emotional Awareness/Regulation” and “Empathy/Perspective Taking” subtasks from the Social Emotional domain.

Figure 9. Comparison of selected subtasks for traditional and balanced short forms



7.6. Performance of short forms

After deciding on the subtasks for inclusion in both short form versions, secondary Graded Response Models were fit to examine the statistical performance of the “balanced” and “traditional” short forms. The Test Information Function of the Full IDELA along with these two short forms are presented in **Figure 10**. As we would expect, eliminating two thirds of the subtasks of the IDELA greatly affects the statistical performance of the tool, as demonstrated by the substantially lower test information functions relative to the full IDELA. Given that the subtasks selected for the “traditional” short form were based exclusively on discrimination parameters, it is not surprising that the Test Information Function for the traditional short form is slightly higher than the balanced IDELA-SF.

Figure 10. Comparison of Test Information Functions

The estimated conditional standard errors of measurement (CSEM) for the full IDELA and short forms were estimated to at the low ($\theta = -2$) at the mean ($\theta = 0$) and the high ($\theta = 2$) level of early learning and development. On the full IDELA, the conditional standard error of measurement for the majority of children in the sample (within two standard deviations of the mean) ranged between 0.214 and 0.306. This indicates that for these children, that the 95% confidence interval for their true score ranges from 0.856 standard deviation units of θ for children with observed scores at the mean to 1.224 standard deviations for children at the 97.5th percentile ($\theta = 2$). In comparison to the Full IDELA, the balanced IDELA-SF generates scores about 38% less precise for children with observed scores at the mean, 72% less precise for children with low scores and 41% less precise for children with high scores. The “traditional” short form is 28% less precise for children with observed scores at the mean, 63% less precise for children with low scores and 37% less precise for children with high scores. In summary, while the balanced IDELA-SF does reasonably well for many children with scores close to the mean, for

children with relatively high and low scores, the difference in the precision between the IDELA-SF and full IDELA grows dramatically.

The two short forms also differ in terms of their conceptual and practical dimensions. For the full IDELA, the average SME score was 2.9 and the average length of administration of the video coding was 26.5 minutes. Both short forms selected subtasks with above-average conceptual ratings. The eight subtasks on the traditionally created short form had an average SME score of 4.0, but the subtasks selected for the balanced IDELA-SF had an even higher average SME score of 6.6. In terms of practicality, the estimated length of traditional short form was over 11 minutes, nearly 25% longer than the eight minutes estimates for the balanced IDELA-SF.

8. Discussion

Direct measures of ECD are a valuable source of information on children's early learning and development, but are expensive and time-consuming to administer. Short forms of assessments like the IDELA could be an attractive means to get some of the benefits of direct assessments at a fraction of the cost and complexity, enabling their use in larger-scale data collections. This paper proposes a three-dimensional human-centered framework that evaluates items or subtasks assessments on their statistical, practical, and conceptual properties to quantify and visualize the tradeoffs to aid a human decision. In addition to item response data on the assessment itself (statistical dimension), this approach requires the collection of information about subtasks or item administration properties (practical dimension) and relevance to the construct being measured by the full tool (conceptual dimension). Creating decision rules and combining these dimensions into human-readable plots allows the selection of subtasks or items for a short tool with improved validity and usability in comparison with a short form created exclusively by examining statistical properties and without the complexity of implementing Automated Test Assembly.

After presenting this framework, the paper attempts to apply it through the current study by collecting data on these three dimensions to create a short form of IDELA. In doing so, two plausible short forms are proposed: a "balanced" approach which considers the practical and conceptual dimensions

of subtasks in addition to their statistical properties and a “traditional” short form based solely on the statistical dimension.

As we would expect, both short forms sacrifice a considerable degree of reliability compared to the full IDELA and result in scores that are about 30-40% less precise at the middle of the distribution. The “traditional” short form does result in scores that are slightly more reliable and precise than those generated from the “balanced” short form. The slight improvement in statistical efficiency of the traditionally created short form comes at a significant cost in terms of its conceptual and practical dimensions. The traditional short form focuses almost exclusively on the Emergent Numeracy (three subtasks) and Emergent Literacy domains (four subtasks) with a single Motor subtask and no Social Emotional subtasks. Just as suggested by the “attenuation paradox,” the subtasks selected for the traditional short form clearly suggest a narrowed construct than the original tool (Boyle, 1991; Little et al., 1999; Loevinger, 1954). As such, scores from this traditional short form may suffer from construct under-representation, and may be less valid at capturing the holistic early learning and development construct IDELA is designed to measure (American Educational Research Association et al., 2014).

In contrast, given the strong preference for a tool that captures the breadth of the full IDELA expressed by subject matter experts, a balanced short form intentionally selected two subtasks from each core domain, ensuring that the construct of early learning and development was adequately represented by the short form. In terms of practicality, the traditional short form selected six subtasks with above-average administration lengths, resulting in an estimated administration time exceeding eleven minutes, about 25% longer than the estimated eight minutes required to administer the balanced short form. In short, the slight reduction in statistical efficiency of the balanced short form in comparison to the traditional short form are more than outweighed by improvements to its conceptual coverage and practical usability.

8.1. Alternative selections of subtasks for a short form

The balanced short form presented in this paper should not be considered the sole correct short form of IDELA. Unlike automated test assembly, this framework does not produce a single short form and ultimately requires a human decision to weigh the tradeoffs of inclusion.

As such, the eight subtasks presented here are a reasonable selection for a general-purpose IDELA-SF that maximizes information for the children in the middle of the distribution. Alternative sets of subtasks could have greater validity for different uses of scores. For example, this analysis ignored the location (difficulty) b parameters, maximizing subtask information regardless of location. This led to scores that were reasonably precise at the middle of the distribution, but substantially less so at the lower end of scores. If a short form wished to maximize precision for children with lower-than-average θ , in order to identify those falling below some threshold, it would require the incorporation of b parameters into the analysis of the statistical dimension and likely lead to the selection of a different set of subtasks. Another possible case would be where the practical demands of data collection necessitated the removal of any subtask requiring special materials. The balanced IDELA-SF presented in this paper is a reasonable approach for a general-purpose short form, but other subsets of subtasks may be more preferred given additional practical limitations or a desire for a different use of scores.

8.2. Alternative approaches to short forms

If constraints for a short form can be operationalized and the mathematical programming does not pose a barrier, ATA is an ideal solution for creating a short form. In addition, multiple matrix sampling is an attractive option when the key demand of a short form is decreasing the length of an individual assessment. Multiple matrix sampling subdivides the full assessment in such a way that no one individual completes each item, but that data on all items are collected, either by creating multiple shorter forms or randomly selecting items for a given test (Shoemaker, 1973). For ECD direct assessments, this would allow users to preserve the entirety of IDELA and the broad construct it measures while still ensuring individual assessments are not too long to administer. This completely addresses concerns with the conceptual dimension of short-forming but only partially addresses the practical concerns. Multiple matrix sampling would still require data collectors to be familiar with all subtask administration rules and possess all required stimulus cards and materials. Training could potentially take even longer than with the full IDELA as data collectors would have to not only learn to administer each subtask with fidelity, but also learn the multiple forms the assessment could take.

8.3. Limitations

The analysis presented in this paper has several limitations and weaknesses related to each of the dimensions of the framework.

8.3.1. Statistical dimension

For the statistical dimension, the sample used to estimate subtask parameters, while large and diverse, is essentially a convenience sample and is not representative of any specific population of interest. As a result, the degree to which the findings about the statistical properties of items are generalizable to new contexts is unknown.

Another limitation of this analysis is that it pools data across 30 countries. Previous research has shown that while IDELA data has a similar factor structure across countries, it does not demonstrate cross-country measurement invariance (Halpin et al., 2019). Differential Item Functioning and Differential test functioning also poses a challenge with internationally used assessments, even those that claim cross-country comparability such as the Programme for International Student Achievement (PISA) (Akour et al., 2015; Hopfenbeck et al., 2018; Rutkowski et al., 2014). A key assumption of all IRT methods is that of conditional independence, that the probability of an individual getting an item correct, conditional on their θ , is independent of any other variable (Yen & Fitzpatrick, 2006). This assumption is violated if another variable is associated with the conditional probability of correctly responding to a question and we can argue that an item may be biased if the cause of DIF is construct-irrelevant. Even when present, DIF effects may be negligible and practically unimportant when DIF effects are weak or DIF direction balances out across the test and results in negligible DTF (Chalmers et al., 2016). However, this calculus may differ when creating short forms. Even if an overall assessment does not display DTF, selecting individual items with DIF could create a biased short form. Creating a short form with a limited number of items could intensify the bias present on the test—future work will consider the degree to which DIF biases short forms.

These sample and representativeness limitations do not detract from the usefulness of the framework but do suggest that, when available, creators of short forms of ECD direct assessments should

focus on using data from their specific population of interest to generate information regarding the statistical properties of the assessment rather than relying on a global dataset and the assumptions that come with it.

8.3.2. Practical and conceptual dimensions

For the practical dimension, as noted earlier, data was quantified from a single, relatively simple, metric of complexity in the length of administration using a small, non-representative, set of video recordings. The videos coded were of IDELA master trainers administering the tool in English, and thus may not be representative of time required to administer IDELA by typical enumerators and in other languages. The level of practicality of subtasks, as suggested above, also encompasses the ease with which data collectors can be trained on the subtask, the ease of scoring the subtask, and any special materials required. These aspects may be difficult to quantify, but a more nuanced approach to this dimension could yield even more information about the practicality of subtasks and their suitability for inclusion in a short form. Users of this framework are encouraged to collect specific data on length of administration from real-world assessments in their context, something that is relatively easy to do by incorporating timestamps into computer-assisted personal interview software.

Finally, the conceptual dimension was considered using a convenience sample of 14 anonymous Subject Matter Experts. These SMEs came from a network representing a range of cultures, nationalities, and experience but this analysis did not collect key demographic or expertise variables that would allow the assessment of the value of their judgement. Similarly, by surveying tool users in English, this may have missed some important constituents such as preschool teachers or policymakers and those that do not speak English. When creating a short form for a specific context, it would be critical to more intentionally reach out to additional relevant stakeholders.

8.3.3. Use of framework beyond IDELA

The IDELA is a popular internationally used measure of ECD but is by no means the only assessment for which this framework could be useful. This framework is most applicable for direct assessments of ECD because of the importance of the practical dimension. This framework would be

particularly well suited to create short forms of other internationally used direct assessments of ECD such as the Anchor Items for Measurement of Early Development (AIM-ECD), the East-Asian Pacific Scales for Early Childhood Development (EAPS-ECD), the Malawi Developmental Assessment Test (MDAT), the Measure of Development and Learning (MODEL), and the Regional Project on Child Development Indicators (PRIDI) (Gladstone et al., 2010; Pushparatnam et al., 2021; Raikes et al., 2019; Rao et al., 2014; Verdisco et al., 2014). Creating short versions of adult-reported measures of ECD could also partially benefit from this framework, but less so given that items in an interview vary less in the practical dimension. Beyond ECD, this framework could be used to aid in the creation of a short form of any direct measure that 1) includes items across multiple dimensions 2) has variation in the administration length and complexity of items 3) has reasonable data on the statistical properties of items.

9. Conclusion

Short forms of ECD assessments could be useful tools in large-scale data collections that improve the quality and quantity of data on developmental outcomes. When creating these short forms, it is important to consider not only the statistical criteria of items, but also their practicality and conceptual relevance. This paper proposes a framework to assess these dimensions of a direct assessment of ECD and applies it to the International Development and Early Learning Assessment to propose a balanced IDELA Short Form that is psychometrically rigorous, easy-to-use, and captures the most important skills within the construct of early learning and development.

10. References

- Akour, M., Sabah, S., & Hammouri, H. (2015). Net and Global Differential Item Functioning in PISA Polytomously Scored Science Items: Application of the Differential Step Functioning Framework. *Journal of Psychoeducational Assessment, 33*(2), 166–176. <https://doi.org/10.1177/0734282914541337>
- Albers, C. A., & Grieve, A. J. (2007). Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development— Third Edition. San Antonio, TX: Harcourt Assessment. *Journal of Psychoeducational Assessment, 25*(2), 180–190. <https://doi.org/10.1177/0734282906297199>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (Joint Committee on Standards for Educational and Psychological Testing (U.S.), Ed.). American Educational Research Association.
- Baker, P. C., & Mott, F. L. (1989). *NLSY Child Handbook 1989: A Guide & Resource Document For The National Longitudinal Survey of Youth 1986 Child Data*. Center for Human Resource Research- The Ohio State University. <https://www.nlsinfo.org/sites/default/files/attachments/121214/ChildHandbook1986%20part%201.pdf>
- Becker, B., Debeer, D., Sachse, K. A., & Weirich, S. (2021). Automated Test Assembly in R: The eatATA Package. *Psych, 3*(2), 96–112. <https://doi.org/10.3390/psych3020010>
- Bennetts, S. K., Mensah, F. K., Westrupp, E. M., Hackworth, N. J., & Reilly, S. (2016). The Agreement between Parent-Reported and Directly Measured Child Language and Parenting Behaviors. *Frontiers in Psychology, 7*. <https://doi.org/10.3389/fpsyg.2016.01710>
- Black, M. M., Walker, S. P., Fernald, L. C. H., Andersen, C. T., DiGirolamo, A. M., Lu, C., McCoy, D. C., Fink, G., Shawar, Y. R., Shiffman, J., Devercelli, A. E., Wodon, Q. T., Vargas-Barón, E., & Grantham-McGregor, S. (2017). Early childhood development coming of age: Science through the life course. *The Lancet, 389*(10064), 77–90. [https://doi.org/10.1016/S0140-6736\(16\)31389-7](https://doi.org/10.1016/S0140-6736(16)31389-7)

- Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*(3), 291–294.
[https://doi.org/10.1016/0191-8869\(91\)90115-R](https://doi.org/10.1016/0191-8869(91)90115-R)
- Bracken, B. (2007). Creating the Optimal Preschool Testing Situation. In B. Bracken & R. Nagle (Eds.), *Psychoeducational Assessment of Preschool Children* (4th ed., pp. 137–154). Routledge.
<https://doi.org/10.4324/9781315089362-9>
- Cavallera, V., Lancaster, G., Gladstone, M., Black, M. M., McCray, G., Nizar, A., Ahmed, S., Dutta, A., Anago, R. K. E., Brentani, A., Jiang, F., Schönbeck, Y., McCoy, D. C., Kariger, P., Weber, A. M., Raikes, A., Waldman, M., van Buuren, S., Kaur, R., ... Janus, M. (2023). Protocol for validation of the Global Scales for Early Development (GSED) for children under 3 years of age in seven countries. *BMJ Open*, *13*(1), e062562. <https://doi.org/10.1136/bmjopen-2022-062562>
- Cella, D., Choi, S. W., Condon, D. M., Schalet, B., Hays, R. D., Rothrock, N. E., Yount, S., Cook, K. F., Gershon, R. C., Amtmann, D., DeWalt, D. A., Pilkonis, P. A., Stone, A. A., Weinfurt, K., & Reeve, B. B. (2019). PROMIS® Adult Health Profiles: Efficient Short-Form Measures of Seven Health Domains. *Value in Health*, *22*(5), 537–544. <https://doi.org/10.1016/j.jval.2019.02.004>
- Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It Might Not Make a Big DIF: Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement*, *76*(1), 114–140. <https://doi.org/10.1177/0013164415584576>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427.
<https://doi.org/10.1037/pas0000626>
- Cromwell, E. A., Dube, Q., Cole, S. R., Chirambo, C., Dow, A. E., Heyderman, R. S., & Van Rie, A. (2014). Validity of US norms for the Bayley Scales of Infant Development-III in Malawian children. *European Journal of Paediatric Neurology*, *18*(2), 223–230.
<https://doi.org/10.1016/j.ejpn.2013.11.011>

- Fernald, L. C. H., & Pitchik, H. O. (2019). The necessity of using direct measures of child development. *The Lancet Global Health*, 7(10), e1300–e1301. [https://doi.org/10.1016/S2214-109X\(19\)30368-7](https://doi.org/10.1016/S2214-109X(19)30368-7)
- Fernald, L. C. H., Prado, E., Kariger, P., & Raikes, A. (2017). A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries (p. 128).
- Frey, B. B. (2018). *The Sage encyclopedia of educational research, measurement, and evaluation*. https://nls.ldls.org.uk/welcome.html?ark:/81055/vdc_100054368665.0x000001
- Gardner, F. (2000). Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review*, 3(3), 185–198. <https://doi.org/10.1023/A:1009503409699>
- Gladstone, M., Lancaster, G. A., Umar, E., Nyirenda, M., Kayira, E., van den Broek, N. R., & Smyth, R. L. (2010). The Malawi Developmental Assessment Tool (MDAT): The Creation, Validation, and Reliability of a Tool to Assess Child Development in Rural African Settings. *PLoS Medicine*, 7(5), e1000273. <https://doi.org/10.1371/journal.pmed.1000273>
- Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., & Strupp, B. (2007). Developmental potential in the first 5 years for children in developing countries. *The Lancet*, 369(9555), 60–70. [https://doi.org/10.1016/S0140-6736\(07\)60032-4](https://doi.org/10.1016/S0140-6736(07)60032-4)
- Halpin, P. F., De Castro, E. F., Petrowski, N., & Cappa, C. (2023). *Monitoring Early Childhood Development at the Population Level: The ECDI2030*. PsyArXiv. <https://doi.org/10.31234/osf.io/6qcjb>
- Halpin, P. F., De Castro, E. F., Petrowski, N., & Cappa, C. (2024). Monitoring early childhood development at the population level: The ECDI2030. *Early Childhood Research Quarterly*, 67, 1–12. <https://doi.org/10.1016/j.ecresq.2023.11.004>
- Halpin, P. F., Wolf, S., Yoshikawa, H., Rojas, N., Kabay, S., Pisani, L., & Dowd, A. J. (2019). Measuring early learning and development across cultures: Invariance of the IDELA across five countries. *Developmental Psychology*, 55(1), 23–37. <https://doi.org/10.1037/dev0000626>

- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, 62(3), 333–353. <https://doi.org/10.1080/00313831.2016.1258726>
- Jeong, J., Franchett, E. E., Ramos de Oliveira, C. V., Rehmani, K., & Yousafzai, A. K. (2021). Parenting interventions to promote early child development in the first three years of life: A global systematic review and meta-analysis. *PLOS Medicine*, 18(5), e1003602. <https://doi.org/10.1371/journal.pmed.1003602>
- John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). Guilford Press.
- Khatib, M. N., Gaidhane, A., Ahmed, M., Saxena, D., & Syed, Z. Q. (2020). Early Childhood Development Programs in Low Middle-Income Countries for Rearing Healthy Children: A Systematic Review. *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*. <https://doi.org/10.7860/JCDR/2020/42134.13445>
- Law, J., & Roy, P. (2008). Parental Report of Infant Language Skills: A Review of the Development and Application of the Communicative Development Inventories. *Child and Adolescent Mental Health*, 13(4), 198–206. <https://doi.org/10.1111/j.1475-3588.2008.00503.x>
- Linden, W. J. van der. (2005). *Linear models of optimal test design*. Springer.
- Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods*, 4(2), 192–211. <https://doi.org/10.1037/1082-989X.4.2.192>

- Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, 51(5), 493–504.
<https://doi.org/10.1037/h0058543>
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. L. Erlbaum Associates.
- Madaan, P., Saini, L., & Sondhi, V. (2021). Development Assessment Scale for Indian Infants: A Systematic Review and Perspective on Dwindling Cutoffs. *Indian Journal of Pediatrics*, 88(9), 918–920. <https://doi.org/10.1007/s12098-021-03671-2>
- McCoy, D. C., Salhi, C., Yoshikawa, H., Black, M., Britto, P. R., & Fink, G. (2018). Home- and center-based learning opportunities for preschoolers in low- and middle-income countries. *Children and Youth Services Review*, 88, 44–56. <https://doi.org/10.1016/j.childyouth.2018.02.021>
- McCoy, D. C., Waldman, M., & Fink, G. (2018). Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported Early Development Instruments. *Early Childhood Research Quarterly*, 45, 58–68. <https://doi.org/10.1016/j.ecresq.2018.05.002>
- Munoz-Chereau, B., Ang, L., Dockrell, J., Outhwaite, L., & Heffernan, C. (2021). Measuring early child development across low and middle-income countries: A systematic review. *Journal of Early Childhood Research*, 19(4), 443–470. <https://doi.org/10.1177/1476718X211020031>
- Pendergast, L. L., Schaefer, B. A., Murray-Kolb, L. E., Svensen, E., Shrestha, R., Rasheed, M. A., Scharf, R. J., Kosek, M., Vasquez, A. O., Maphula, A., Costa, H., Rasmussen, Z. A., Yousafzai, A., Tofail, F., Seidman, J. C., & The MAL-ED Network Investigators. (2018). Assessing development across cultures: Invariance of the Bayley-III Scales Across Seven International MAL-ED sites. *School Psychology Quarterly*, 33(4), 604–614.
<https://doi.org/10.1037/spq0000264>
- Pisani, L., Borisova, I., & Dowd, A. J. (2015). *International Development and Early Learning Assessment Technical Working Paper* (p. 26). Save the Children. <https://idela-network.org/resource/international-development-and-early-learning-assessment-technical-working-paper/>

- Pisani, L., Borisova, I., & Dowd, A. J. (2018). Developing and validating the International Development and Early Learning Assessment (IDELA). *International Journal of Educational Research, 91*, 1–15. <https://doi.org/10.1016/j.ijer.2018.06.007>
- Pushparatnam, A., Luna Bazaldua, D. A., Holla, A., Azevedo, J. P., Clarke, M., & Devercelli, A. (2021). Measuring Early Childhood Development Among 4–6 Year Olds: The Identification of Psychometrically Robust Items Across Diverse Contexts. *Frontiers in Public Health, 9*, 569448. <https://doi.org/10.3389/fpubh.2021.569448>
- R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.2.1 (Bird Hippie)) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Raikes, A., Koziol, N., Janus, M., Platas, L., Weatherholt, T., Smeby, A., & Sayre, R. (2019). Examination of school readiness constructs in Tanzania: Psychometric evaluation of the MELQO scales. *Journal of Applied Developmental Psychology, 62*, 122–134. <https://doi.org/10.1016/j.appdev.2019.02.003>
- Rao, N., Chan, S. W. Y., Su, Y., Richards, B., Cappa, C., De Castro, E. F., & Petrowski, N. (2021). Measuring Being “Developmentally on Track”: Comparing Direct Assessment and Caregiver Report of Early Childhood Development in Bangladesh, China, India and Myanmar. *Early Education and Development, 1–23*. <https://doi.org/10.1080/10409289.2021.1928446>
- Rao, N., Sun, J., Ng, M., Becher, Y., Lee, D., Ip, P., & Bacon-Shone, J. (2014). *Validation, Finalization and Adoption of the East Asia-Pacific Early Child Development Scales (EAP-ECDS)*. UNICEF, East and Pacific Regional Office. <https://arnec.net/sites/default/files/2022-09/EAP-ECDS-Final-Report1.pdf>
- Renk, K. (2005). Cross-Informant Ratings of the Behavior of Children and Adolescents: The “Gold Standard.” *Journal of Child and Family Studies, 14*(4), 457–468. <https://doi.org/10.1007/s10826-005-7182-2>

- Rubio-Codina, M., Araujo, M. C., Attanasio, O., Muñoz, P., & Grantham-McGregor, S. (2016). Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies. *PLOS ONE*, *11*(8), e0160962.
<https://doi.org/10.1371/journal.pone.0160962>
- Rutkowski, L., Davier, M. von, & Rutkowski, D. (Eds.). (2014). Handbook of International large-scale assessment: Background, technical issues, and methods of data analysis. CRC Press, Taylor & Francis Group.
- Sabanathan, S., Wills, B., & Gladstone, M. (2015). Child development assessment tools in low-income and middle-income countries: How can we use them more appropriately? *Archives of Disease in Childhood*, *100*(5), 482–488. <https://doi.org/10.1136/archdischild-2014-308114>
- Sachse, S., & Suchodoletz, W. V. (2008). Early Identification of Language Delay by Direct Language Assessment or Parent Report? *Journal of Developmental & Behavioral Pediatrics*, *29*(1), 34–41.
<https://doi.org/10.1097/DBP.0b013e318146902a>
- Salsman, J. M., Schalet, B. D., Park, C. L., George, L., Steger, M. F., Hahn, E. A., Snyder, M. A., & Cella, D. (2020). Assessing meaning & purpose in life: Development and validation of an item bank and short forms for the NIH PROMIS®. *Quality of Life Research*, *29*(8), 2299–2310.
<https://doi.org/10.1007/s11136-020-02489-3>
- Samejima, F. (1968). ESTIMATION OF LATENT ABILITY USING A RESPONSE PATTERN OF GRADED SCORES¹. *ETS Research Bulletin Series*, *1968*(1), i–169.
<https://doi.org/10.1002/j.2333-8504.1968.tb00153.x>
- Save the Children. (2019). *International Development and Early Learning Assessment (IDELA): Administration Guide*. https://idela-network.org/wp-content/uploads/2019/04/Administration-Guide-and-IDELA-Tool_2019.pdf
- Save the Children. (2024). *About IDELA: Countries of Implementation*. International Development & Early Learning Assesment. <https://idela-network.org/about/countries-where-idela-is-in-use/>
- Shoemaker, D. M. (1973). Principles and procedures of multiple matrix sampling. Ballinger Pub. Co.

- Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111. <https://doi.org/10.1037/1040-3590.12.1.102>
- StataCorp. (2019). *Stata Statistical Software* (Version Release 16) [Computer software]. StataCorp LLC.
- van Buuren, S., & Eekhout, I. (2024). *Child development with the D-score*. CRC Press, Taylor et Francis Group.
- Verdisco, A., Cueto, C., Santiago, Thompson, J., & Neuschmidt, O. (2014). *Urgency and Possibility: First Initiative of Comparative Data on Child Development in Latin America*. Inter-American Development Bank. <https://publications.iadb.org/publications/english/document/PRIDI-Urgency-and-Possibility.pdf>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. <https://doi.org/10.21105/joss.01686>
- Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, M. B. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis: An introduction for psychologists*. (pp. 39–61). American Psychological Association. <https://doi.org/10.1037/12350-003>
- Wilcox, R. (2017). *Introduction to Robust Estimation and Hypothesis Testing (Fourth Edition)*. Elsevier. <https://doi.org/10.1016/B978-0-12-804733-0.00014-7>
- Wolf, S., Halpin, P., Yoshikawa, H., Dowd, A. J., Pisani, L., & Borisova, I. (2017). Measuring school readiness globally: Assessing the construct validity and measurement invariance of the International Development and Early Learning Assessment (IDELA) in Ethiopia. *Early Childhood Research Quarterly, 41*, 21–36. <https://doi.org/10.1016/j.ecresq.2017.05.001>
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational measurement* (4. ed). American Council on Education [u.a.].

11. Supplemental Materials Annex A: IDELA User Survey

IDELA User Survey for Creating Short Form

Thank you for considering participation in this survey! Please read the below and indicate if you wish to participate.

Study Title: *A Framework for Creating Short Forms of Internationally Used Direct Assessment of Early Childhood Development*

Researcher: [REDACTED]

Faculty Advisor: [REDACTED]

Version Date: 20-October-2021

What is the purpose of this research?

When creating short forms of Early Childhood Development assessments, it's important to consider not only the statistical properties of the questions on the assessment, but also the practical considerations of the assessment and the "face validity" of subtasks that comprise a short form.

This research aims to develop a Short Form of the International Development and Early Learning Assessment (IDELA). We are surveying experts like you to understand their opinions around the subtasks on the IDELA and the uses and properties of a potential short form. This information will be used to help ensure a Short Form IDELA is relevant to users.

What can I expect if I take part in this research?

This research involves responding to a short survey about your opinion of the subtasks on IDELA and what you value in a short form of the IDELA.

What should I know about a research study?

- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
- You can ask all the questions you want before you decide.

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, talk to [REDACTED] at [REDACTED]

Do you agree to participate in this survey?

Yes

No

Background

How many separate data collections have you been involved with working on IDELA?

- None
- One
- 2-4
- 5-10
- More than 10

Have you ever led an IDELA training yourself?

- Yes
- No

On average, how long would you estimate that the full IDELA has taken to administer **in minutes** when you have used it?

Subtask Ranking

Now we are going to ask you about different subtasks on the IDELA and which subtasks you feel like best capture children's development in the different domains of IDELA. For each domain, we ask you identify the **MOST** and **2nd MOST** important subtasks for measuring the domain, as well as the **LEAST** important subtask. You may find it useful to have a copy of the IDELA tool at hand to refer to while completing this section

Within the **Emergent Numeracy** domain, please indicate which subtasks are MOST and LEAST important to measuring Emergent Numeracy skills.

MOST important *

- | | | |
|---|--|---|
| <input type="radio"/> Comparison by Size and Length | <input type="radio"/> Number Identification | <input type="radio"/> Shape Identification |
| <input type="radio"/> Puzzle Completion | <input type="radio"/> Sorting and Classification | <input type="radio"/> One-to-one Correspondence |
| <input type="radio"/> Addition and Subtraction | | |

2nd MOST important *

- | | | |
|---|--|---|
| <input type="radio"/> Comparison by Size and Length | <input type="radio"/> Number Identification | <input type="radio"/> Shape Identification |
| <input type="radio"/> Puzzle Completion | <input type="radio"/> Sorting and Classification | <input type="radio"/> One-to-one Correspondence |
| <input type="radio"/> Addition and Subtraction | | |

LEAST important *

- | | | |
|---|--|---|
| <input type="radio"/> Comparison by Size and Length | <input type="radio"/> Number Identification | <input type="radio"/> Shape Identification |
| <input type="radio"/> Puzzle Completion | <input type="radio"/> Sorting and Classification | <input type="radio"/> One-to-one Correspondence |
| <input type="radio"/> Addition and Subtraction | | |

Within the **Emergent Literacy** domain, please indicate which subtasks are MOST and LEAST important to measuring Emergent Literacy skills.

MOST important *

- | | | |
|--|---|---|
| <input type="radio"/> Print Awareness | <input type="radio"/> Expressive Vocabulary | <input type="radio"/> Letter Identification |
| <input type="radio"/> Emergent Writing | <input type="radio"/> First Letter Sounds | <input type="radio"/> Oral Comprehension |

2nd MOST important *

Print Awareness Expressive Vocabulary Letter Identification
 Emergent Writing First Letter Sounds Oral Comprehension

LEAST important *

Print Awareness Expressive Vocabulary Letter Identification
 Emergent Writing First Letter Sounds Oral Comprehension

Within the **Social-Emotional** domain, please indicate which subtasks are MOST and LEAST important to measuring Social-Emotional skills.

MOST important *

Friends Emotional Awareness / Regulation Empathy / Perspective Taking
 Personal Knowledge/Self-Awareness Conflict Resolution

2nd MOST important *

Friends Emotional Awareness / Regulation Empathy / Perspective Taking
 Personal Knowledge/Self-Awareness Conflict Resolution

LEAST important *

Friends Emotional Awareness / Regulation Empathy / Perspective Taking
 Personal Knowledge/Self-Awareness Conflict Resolution

Within the **Motor** domain, please indicate which subtasks are MOST and LEAST important to measuring Motor skills.

MOST important *

Hopping Copying a Shape Drawing a Person
 Folding Paper

2nd MOST important *

Hopping Copying a Shape Drawing a Person
 Folding Paper

LEAST important *

Hopping Copying a Shape Drawing a Person
 Folding Paper

Within the **Executive Function** domain, which subtask do you believe is MOST important to measuring Executive Function skills?

- Short-term Memory
- Inhibitory Control (Head, Toes, Knees, Shoulders)

Conclusion

This is the last section. In this section, please indicate how strongly you **AGREE** or **DISAGREE** with the following statements.

Read the following statements Strongly Disagree Disagree Agree Strongly Agree

It's very important that the Short Form takes less than 10 minutes to complete.

I won't use a Short Form if it doesn't include my favorite subtasks.

I am excited about using a Short Form IDELA.

It's very important that a Short Form IDELA includes subtasks from ALL domains of the assessment.

It's very important for a Short Form IDELA to have strong psychometric properties, even if it means it some domains are not included.

If a Short Form IDELA were available today, how would you use it?

Do you have any other feedback or advice as we create a Short Form IDELA?

If you would like us to follow up with you, please indicate your contact information.

12. Supplemental Materials Annex B: IDELA Subtask Descriptives and Correlations

Figure B1

Distribution of Scores on IDELA Subtasks

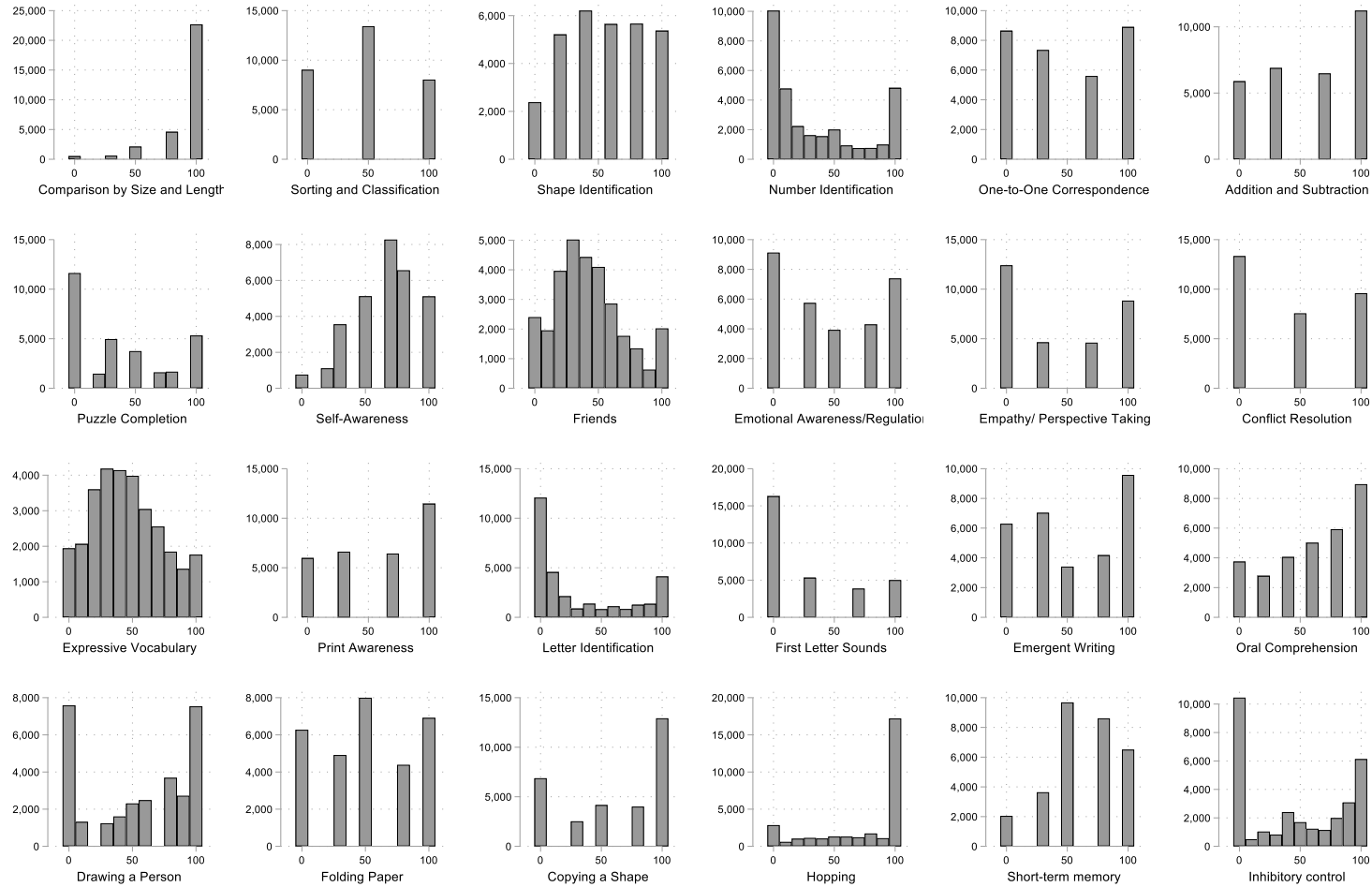


Figure B2

Heat Map of Pairwise Correlations for All IDELA Subtasks

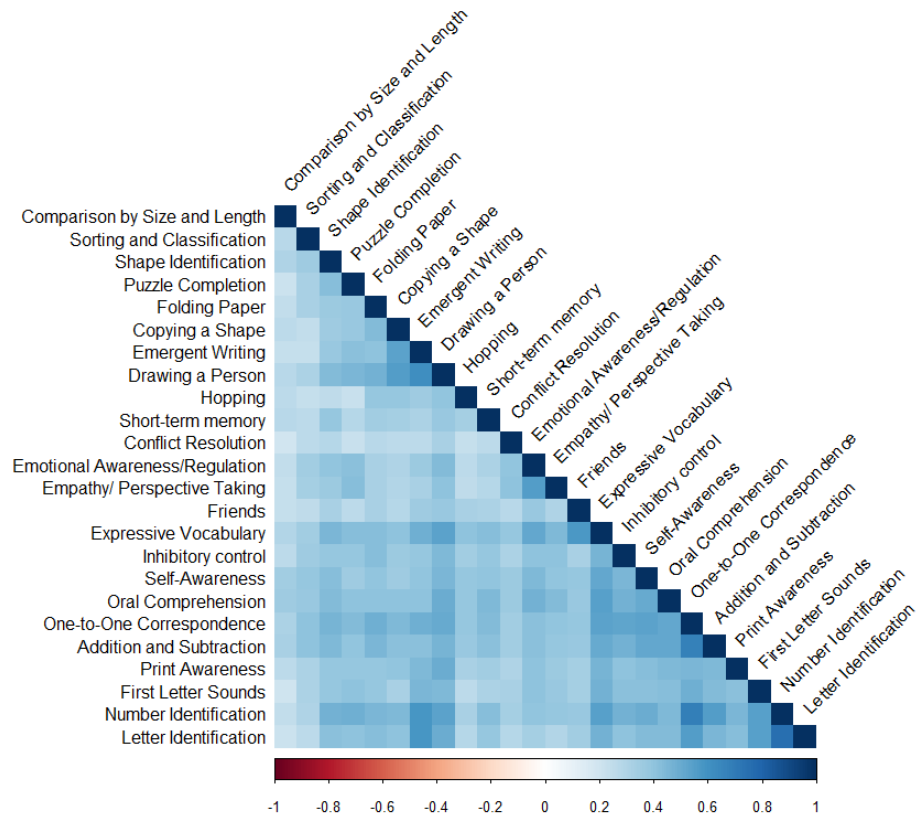


Table B1

Correlation Matrix of IDELA Subtasks

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	
(1) Comparison by Size and Length	1.00																								
(2) Sorting and Classification	0.27	1.00																							
(3) Shape Identification	0.30	0.36	1.00																						
(4) Number Identification	0.24	0.31	0.48	1.00																					
(5) One-to-One Correspondence	0.31	0.40	0.47	0.68	1.00																				
(6) Addition and Subtraction	0.33	0.40	0.45	0.55	0.67	1.00																			
(7) Puzzle Completion	0.21	0.33	0.43	0.48	0.44	0.39	1.00																		
(8) Self-Awareness	0.35	0.38	0.42	0.49	0.54	0.51	0.36	1.00																	
(9) Friends	0.23	0.26	0.32	0.38	0.38	0.37	0.26	0.38	1.00																
(10) Emotional Awareness/Regulation	0.25	0.34	0.40	0.40	0.41	0.41	0.41	0.44	0.37	1.00															
(11) Empathy/ Perspective Taking	0.24	0.34	0.36	0.38	0.40	0.38	0.43	0.39	0.30	0.55	1.00														
(12) Conflict Resolution	0.19	0.27	0.28	0.33	0.34	0.34	0.23	0.36	0.28	0.39	0.40	1.00													
(13) Expressive Vocabulary	0.29	0.35	0.46	0.54	0.54	0.50	0.43	0.52	0.58	0.51	0.44	0.38	1.00												
(14) Print Awareness	0.27	0.31	0.39	0.46	0.45	0.44	0.38	0.43	0.34	0.40	0.36	0.31	0.47	1.00											
(15) Letter Identification	0.22	0.26	0.41	0.76	0.55	0.45	0.40	0.43	0.34	0.33	0.29	0.28	0.47	0.43	1.00										
(16) First Letter Sounds	0.21	0.32	0.39	0.55	0.49	0.43	0.41	0.41	0.33	0.40	0.38	0.33	0.47	0.42	0.54	1.00									
(17) Emergent Writing	0.23	0.23	0.38	0.59	0.48	0.41	0.41	0.39	0.35	0.36	0.32	0.26	0.48	0.45	0.58	0.45	1.00								
(18) Oral Comprehension	0.35	0.38	0.44	0.45	0.51	0.51	0.40	0.51	0.37	0.48	0.43	0.36	0.54	0.45	0.43	0.43	0.40	1.00							
(19) Drawing a Person	0.28	0.31	0.44	0.53	0.50	0.46	0.46	0.45	0.38	0.43	0.40	0.32	0.53	0.49	0.50	0.44	0.61	0.49	1.00						
(20) Folding Paper	0.24	0.33	0.36	0.45	0.48	0.45	0.37	0.40	0.32	0.33	0.33	0.28	0.43	0.39	0.43	0.39	0.41	0.41	0.47	1.00					
(21) Copying a Shape	0.26	0.24	0.36	0.44	0.44	0.41	0.37	0.36	0.27	0.30	0.29	0.26	0.40	0.40	0.40	0.32	0.53	0.40	0.55	0.43	1.00				
(22) Hopping	0.27	0.24	0.25	0.32	0.38	0.37	0.22	0.38	0.32	0.27	0.26	0.24	0.41	0.32	0.28	0.27	0.35	0.39	0.39	0.38	0.38	1.00			
(23) Short-term memory	0.28	0.26	0.38	0.42	0.43	0.41	0.28	0.40	0.31	0.31	0.28	0.27	0.43	0.35	0.39	0.32	0.32	0.45	0.38	0.34	0.33	0.33	1.00		
(24) Inhibitory control	0.26	0.36	0.38	0.47	0.52	0.48	0.39	0.45	0.33	0.40	0.40	0.31	0.47	0.40	0.40	0.42	0.38	0.48	0.44	0.43	0.36	0.35	0.39	1.00	

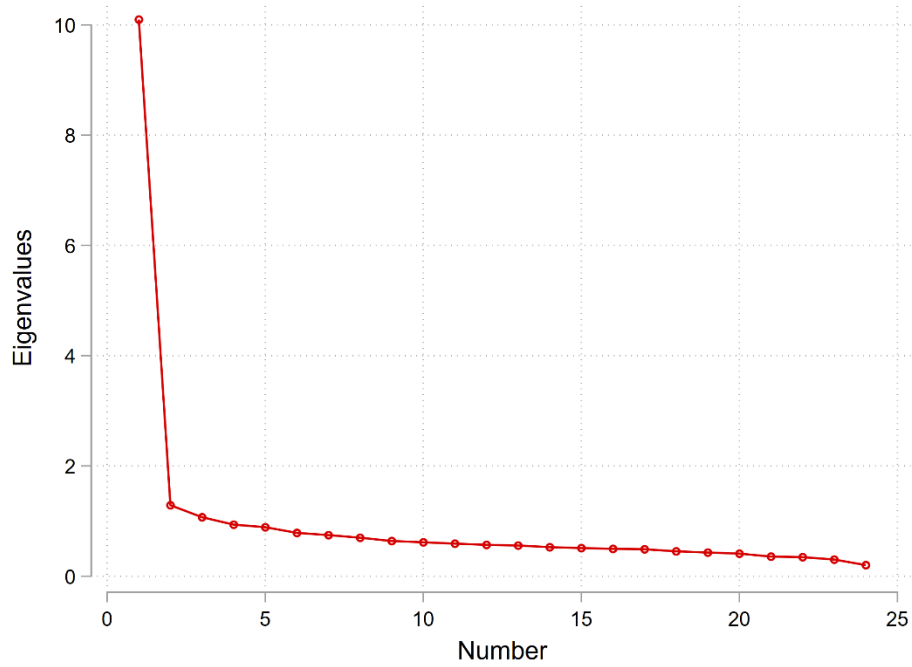
13. Supplemental Materials Annex C: IDELA Dimensionality and Internal Consistency*Figure C1**Scree Plot of Eigenvalues After Factor Analysis of 24 IDELA Subtasks*

Table C1

Cronbach's Alpha of Full IDELA subtasks

Subtask	Sign	Item-test correlation	Item-rest correlation	Alpha
Comparison by Size and Length	+	0.4334	0.4021	0.9367
Sorting and Classification	+	0.5337	0.4824	0.936
Shape Identification	+	0.6345	0.5989	0.9343
Number Identification	+	0.7642	0.7337	0.9322
One-to-One Correspondence	+	0.7695	0.7376	0.932
Addition and Subtraction	+	0.7228	0.6872	0.9329
Puzzle Completion	+	0.6212	0.5765	0.9346
Self-Awareness	+	0.6832	0.658	0.9341
Friends	+	0.5522	0.5173	0.9354
Emotional Awareness/Regulation	+	0.6454	0.6009	0.9342
Empathy/ Perspective Taking	+	0.6158	0.5644	0.9349
Conflict Resolution	+	0.5284	0.4686	0.9366
Expressive Vocabulary	+	0.7418	0.7185	0.9332
Print Awareness	+	0.6564	0.6139	0.934
Letter Identification	+	0.6892	0.6506	0.9334
First Letter Sounds	+	0.6613	0.6193	0.9339
Emergent Writing	+	0.68	0.6391	0.9336
Oral Comprehension	+	0.7103	0.6774	0.9331
Drawing a Person	+	0.7354	0.7002	0.9326
Folding Paper	+	0.6397	0.599	0.9342
Copying a Shape	+	0.6224	0.5745	0.9347
Hopping	+	0.5365	0.4892	0.9357
Short-term memory	+	0.5754	0.5392	0.9351
Inhibitory control	+	0.672	0.6283	0.9338
Hopping	+	0.4334	0.4021	0.9367
Test scale				0.9368

14. Supplemental Materials Annex D: IDELA Graded Response Model Parameter Estimates

Table D1: Full Parameter Estimates for IDELA Graded Response Model (Panel 1)

		Comparison by Size and Length	Sorting and Classification	Shape Identification	Number Identification	One-to-One Correspondence	Addition and Subtraction	Puzzle Completion	Self-Awareness	Friends	Emotional Awareness/Regulation	Empathy/Perspective Taking	Conflict Resolution
Location (b) parameters	a parameter	1.170	1.140	1.419	2.483	2.485	2.099	1.420	1.772	1.207	1.491	1.382	1.112
	Number of b parameters	4	2	5	20	3	3	12	6	10	4	3	2
	>=5				-0.566								
	>=10				-0.291					-2.510			
	>=13												
	>=15				-0.097								
	>=17												
	>=20			-2.242	0.043			-0.462	-2.815		-1.851		
	>=25	-4.071			0.135			-0.293			-0.827		
	>=30				0.223			-0.288					
	>=33					-0.711	-1.144	-0.066	-2.170			-0.402	
	>=35				0.312								
	>=38												
	>=40			-1.047	0.399			0.346			-0.273		
	>=45				0.489								
	>=50	-3.305	-0.937		0.575			0.351	-1.279	0.339	-0.075		-0.312
	>=55				0.734								
	>=60			-0.184	0.799			0.872		0.956			
	>=63												
	>=65				0.860								
>=67					0.055	-0.292	0.880	-0.563			0.192		
>=70				0.921					1.491				
>=75	-2.195			0.969			1.141			0.423			
>=80			0.554	1.028			1.334		1.923				
>=83							1.336	0.392					
>=85				1.086									
>=88													
>=90				1.151					2.368				
>=95				1.249									
>=100	-1.121	1.124	1.473	1.398	0.667	0.417	1.455	1.329	2.650	1.036	0.837	0.833	

Note: Subtasks are colored according to domain with magenta (Emergent Numeracy), blue (Social Emotional), Emergent Literacy (yellow), Motor (green), and Executive Function (purple). Location (b) parameters refer to the percent correct level on the subtask. Subtasks have different b parameters depending on the total possible points on the subtask.

Full Parameter Estimates for IDELA Graded Response Model (Panel 2)

		Expressive Vocabulary	Print Awareness	Letter Identification	First Letter Sounds	Emergent Writing	Oral Comprehension	Drawing a Person	Folding Paper	Copying a Shape	Hopping	Short-term memory	Inhibitory control
Location (b) parameters	a parameter	1.991	1.631	2.068	1.824	1.654	1.895	2.012	1.482	1.399	1.202	1.311	1.569
	Number of b parameters	20	3	20	3	4	5	8	4	4	10	4	10
	>=5	-2.065		-0.354									
	>=10	-1.813		-0.087							-2.339		-0.594
	>=13							-0.940					
	>=15	-1.506		0.116									
	>=17												
	>=20	-1.186		0.248			-1.603				-2.123		-0.528
	>=25	-0.912		0.342		-1.257		-0.772	-1.259	-1.197		-2.521	
	>=30	-0.632		0.389							-1.815		-0.389
	>=33		-1.250		0.071								
	>=35	-0.395		0.440									
	>=38							-0.619					
	>=40	-0.159		0.494				-1.077				-1.535	
	>=45	0.060		0.603									
	>=50	0.283		0.655		-0.350		-0.427	-0.537	-0.767	-1.308	-1.418	0.021
	>=55	0.506		0.707									
	>=60	0.709		0.751				-0.524			-1.060		0.232
	>=63								-0.161				
	>=65	0.889		0.856									
>=67		-0.337		0.697									
>=70	1.091		0.911							-0.832		0.383	
>=75	1.282		0.978			0.081		0.123	0.485	-0.195	0.081		
>=80	1.478		1.050				0.046				-0.639		
>=83													
>=85	1.663		1.184										
>=88								0.550					
>=90	1.855		1.298								-0.387		
>=95	2.081		1.444									0.800	
>=100	2.251	0.446	1.669	1.320	0.643	0.721	0.892	1.140	0.298	-0.238	1.349	1.268	