Creation and validation of the LEVANTE core tasks: Internationalized measures of learning and development for children ages 5-12 years

[1] & Michael C. Frank[1]

[1] Stanford University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. : Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Michael C. Frank: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to , Building 420, 450 Jane Stanford Way, Stanford, CA 94305. E-mail: mcfrank@stanford.edu

Abstract

We present the Learning Variability Network Exchange (LEVANTE) core tasks, a set of nine short and engaging computer adaptive tasks designed to assess learning and development in children ages 5–12 years across a wide range of languages and cultures. Using a simple and uniform multi-alternative forced choice format, these tasks measure constructs including math, execxutive function, reasoning, and social cognition and can be administered on a tablet or computer both in person or remotely. We describe the design and selection of these instruments, and then report on their reliability and validity in a pilot sample of XYZ children recruited in Colombia, Germany, and Canada. Tasks are scored using item response theory models. These models can be used to create computer adaptive versions of the tasks, allowing the entire battery to be given in under an hour. We discuss the use and extension of these tasks in the service of creating an open dataset to describe variability in children's development and learning across contexts.

Keywords: cognitive development

Word count: X

Creation and validation of the LEVANTE core tasks: Internationalized measures of learning and development for children ages 5-12 years

Introduction

Developmental variability and change during childhood is a focus of intense theoretical and practical interest. From tracking children's growth over time to evaluating intervention outcomes or exploring environmental and contextual moderators, a wide range of scientific goals require accurate assessments. Ideal psychological measures provide efficient, reliable, and valid measures of particular constructs that can be applied across a range of ages, situations, and contexts. Yet, in most cases, a large gap separates the situation of a researcher searching for measures from this ideal.

Because children's overall capacities are so dependent on their age, it can be very challenging to use the same measure across children of different ages. Young children require simple tasks that are not verbally demanding, while older children can answer more complicated questions. In addition, younger children typically require shorter tasks, often reducing measurement reliability. Yet giving different tasks to different ages can mean that scores are not comparable to one another, making tracking developmental growth challenging in many domains.

A second set of challenges concern cross-context comparisons. Ideal developmental measures should be validated in a global context and applicable to children across many cultures and languages. Child development is an issue of global importance [LANCET CITES], yet the vast majority of measures are developed in very specific (often English-speaking) contexts. Providing cross-culturally validated measures allows for the collection of comparable data across contexts, opening up opportunities for theoretical synthesis.

A final set of challenges has to do with accessibility. Many gold-standard measures are commercially distributed. They are costly for researchers to use, and in addition, publishers may place barriers on new translation and adaptation. Publishers also typically hold both item information and normative data closely, blocking many types of secondary investigation.

In the context of these challenges, we describe the Learning Variability Network Exchange (LEVANTE) (**frank2025?**). LEVANTE provides a technical framework for data collection: researchers can use the LEVANTE dashboard to assign both surveys and tasks to children, caregivers, and teachers. Data collected via the dashboard are harmonized and validated and become accessible through a data repository, first to the researchers who collect them and eventually – through regular releases – to the broader research community. Through a partnership with the Jacobs Foundation, sites around the world are funded to collect longitudinal data from children using the LEVANTE framework. The eventual goal of LEVANTE is to create a large dataset documenting children's learning and development across contexts.

The current manuscript introduces the LEVANTE core tasks, the behavioral measures for children developed for LEVANTE. In our initial development of the framework, we cast a broad net for important constructs in child development with well-accepted measures that had been used internationally. This process is described in (**frank2025?**). The broad constructs that we selected were executive function, language, mathematics, reasoning, and social cognition, with these being instantiated through a number of well-accepted tasks.

To create our core tasks, we selected pre-existing measures from the literature that tapped each of these constructs, when possible prioritizing measures with strong psychometric properties, previous use across a broad range of cultures, applicability across a broad range of ages, and lack of commercial or licensing constraints. This process yielded a series of measures, which we implemented in an open source web platform. Table 1 shows these tasks, organized by construct.

In addition to the constructs described above, we were also interested in the assessment of literacy. The LEVANTE core tasks battery makes use of a number of previously-validated literacy tasks from the Rapid Online Assessment of Reading (Yeatman et al., 2021) , including single word reading, sentence reading efficiency, and phonological awareness. We do not report on these tasks here, though we make use of them for validation of language measures. Similarly, we included a commercially-available broad measure of executive function, the Minnesota Executive Function Scale (MEFS) (**mefs?**) for validation purposes.

Here we report on the development and validation of these tasks. This is an iterative process in which data from 5–12 year old children has been collected across three sites: Bogota, Colombia; Leipzig, Germany; and Ottawa, Canada. In some cases that we note below, we used these data during the data collection process to make minor changes to the tasks. We use data from these three pilot sites both to provide initial evidence on the reliability and validity of the measures and to develop efficient, computer adaptive (CAT) versions of nearly all of the tasks.

A key component of this process is the use of psychometric models based on item-response theory (IRT) (**embretson2001?**). IRT models provide a family of models that allow the joint estimation of the difficulty of individual task items (e.g., math questions) and the ability of individual children. A fitted IRT model provides task parameters that can be used to estimate the ability of a new test taker given their responses on some or all of the same items. In addition, IRT parameters are used in the construction of CATs, which choose the most relevant items to give to estimate the ability of a particular individual. Critically for our purposes, the use of IRT models means that we can provide comparable scores on the same scale to a younger child who saw mostly easier task items and an older child who saw harder items; these models thus allow us to address our first key challenge posed above.

Because our data come from three sites, each with their own translations and adaptations of the specific tasks, we can also use multi-group IRT models to explore the question of invariance: whether measures function similarly across different groups (**bornstein2016?**). While measurement invariance is more commonly discussed in the factor analytic literature, it is also applicable to IRT (where it is sometimes analyzed at the level of individual test items as "differential item function" across groups) (**thissen2024?**). Here we use multigroup model comparisons (described below) to investigate whether our tasks measure similarly structured constructs across groups. In particular, where possible, we aim for *scalar invariance*, in which individuals from different groups still show the same relative ordering of difficulty across items (e.g., they still find fractions items harder than division items in a math test). In some cases, we may fall back to *metric invariance*, in which items show different difficulties across groups, or *configural invariance*, in which items show different degrees of ability discrimination as well (e.g., if some problem types are unfamiliar to children in one group and so do not discriminate between high and low ability children). These models allow us to begin to address the second challenge posed above.

In what follows, we begin by describing the nine LEVANTE core tasks, organized by construct. We then discuss the process of translation and adaptation that produced the Spanish and German versions of these tasks from the original English source. We then discuss our pilot data collection efforts in Colombia, Germany, and Canada. We present our IRT-based scoring techniques and the results of multi-group comparison. Using scores from these analyses, we then present preliminary evidence on the reliability and validity of the tasks, recognizing that in many cases these tasks are still under construct and we anticipate increases in reliability as we iteratively improve items.

We end by discussing future plans for further internationalization and downward extension of the tasks. Critically, LEVANTE embraces open science values, aiming to create measures and data that are permissively licensed and reusable and extensible by the international research community. These values address our final challenge posed above: the aim of LEVANTE is to minimize barriers to reuse, accelerating progress towards a global science of learning and development.

The LEVANTE core tasks

The LEVANTE core tasks are implemented using jsPsych (**deleeuw?**) and can be presented in a web browser on a tablet or laptop, with responses possible using a touchscreen, keyboard, or mouse. Because of this variability in format of administration, they focus on response correctness not reaction time and so they are mostly untimed. The tasks are designed for simplicity and clarity so as to be accessible to children across a wide age range, and so with only modest exceptions, nearly all are in the format of a multi-alternative forced choice with a maximum of four choices. This uniformity of format means that in most cases instructions can be short and easy to understand, minimizing delays when the tasks are given in sequence as a battery. Figure @ref{fig:tasks} shows screenshots from a number of tasks. In the remainder of this section, we briefly present each of the LEVANTE core tasks.

Language

Sentence Understanding. The Test for Reception of Grammar (TROG) (**bishop1982?**) is a multiple-choice measure of receptive grammatical understanding. On each trial, the child hears a spoken sentence and is asked to select one of four pictures that best matches its meaning. The original test contained 20 blocks, each with four items assessing the same grammatical structure. In our adaptation (based on the original TROG, which was permissively licensed for reuse), we removed a small number of items due to changes in cultural norms. In addition, based on early pilot testing showing that many trials were easy for older children, we added a set of several dozen more challenging sentences. All illustrations were remade with details intended to be accessible across a broad range of cultures.

Vocabulary. The Vocabulary task was developed as a non-commercial, open alternative to tasks such as the Peabody Picture Vocabulary (**peabody?**) and the NIH

Toolbox Picture Vocabulary Task (**nihtoolbox?**).

Math

Reasoning

Matrix Reasoning.

Shape Rotation.

Executive Function

Same Difference Selection. The Same Different Selection task (Obradovi'c & Sulik, year?) is designed to assess cognitive flexibility in children. It draws upon elements from the 'Something's the Same' task (Willoughby et al., 2012) and the 'Flexible Item Selection Task' (Jacques & Zelazo, 2001). In this task, children are presented with sets of items that vary along multiple dimensions, such as shape, color, size, and number. They are required to identify similarities and differences between items based on these dimensions, thereby engaging their ability to shift attention and adapt to changing rules or criteria.

Hearts and Flowers. The Hearts and Flowers task (Davidson et al., 2006) assesses inhibitory control and cognitive flexibility in individuals aged 3.5 years to adulthood. Participants respond according to stimulus type: pressing a key on the same side as a heart (congruent rule), and on the opposite side for a flower (incongruent rule). The task includes three blocks - congruent (hearts only), incongruent (flowers only), and mixed (hearts and flowers). The congruent block serves as a baseline with minimal executive demands. Inhibitory control is typically measured via performance on the incongruent block, which requires overriding a spatially dominant response, while cognitive flexibility is assessed using the mixed block, which demands switching between rules based on the stimulus (Wright & Diamond, 2014).

Memory Game. The Corsi Block task is a widely used measure of visuospatial short-term memory (**corsi1972?**). In the standard version, a set of four blocks is arranged in a fixed spatial configuration. During each trial, a subset of blocks lights up one at a time in a specific sequence. The child is required to reproduce the sequence by clicking the blocks in the same order. The task begins with short sequences (e.g., two items) and gradually increases in difficulty (up to five or more) until the child fails two sequences of the same length. The longest correctly reproduced sequence reflects their visuospatial span. When adapted for younger children or digital administration, the number of visible blocks may be reduced (e.g., four blocks), and span lengths are typically capped at five to reduce task complexity.

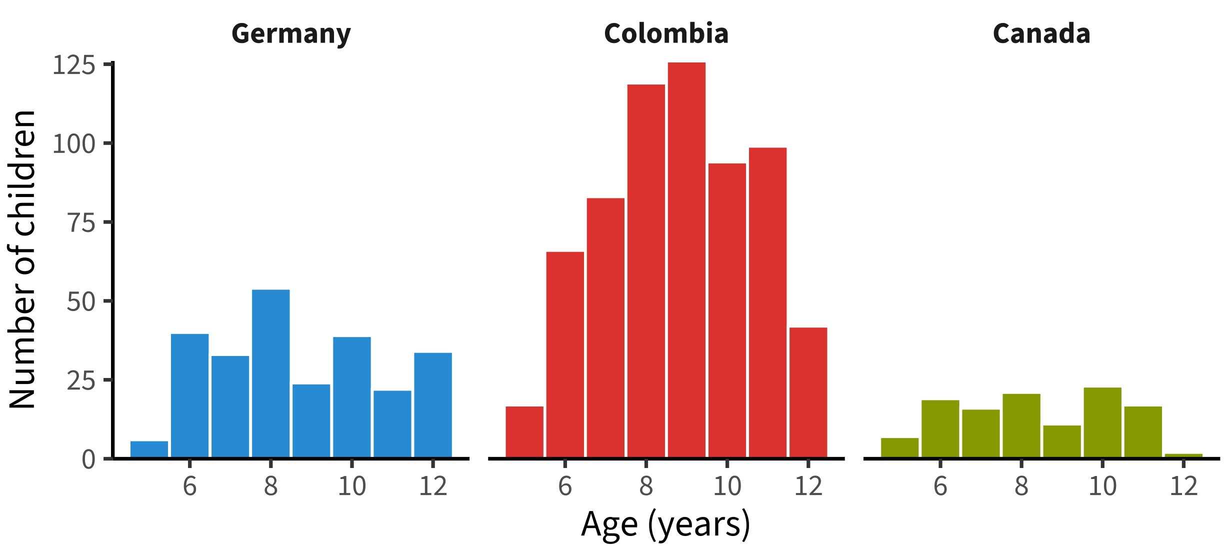
Social Cognition

Stories task.

Translation and Adaptation

Need some discussion here of how translation was done and checked by sites

Pilot Data Collection



Colombia

Germany

Canada

Scoring

IRT Calibration and Model Selection

Multigroup Calibration and Measurement Invariance

Model comparisons employed likelihood ratio test (?), changes in Bayesian Information Criterion (Δ BIC).

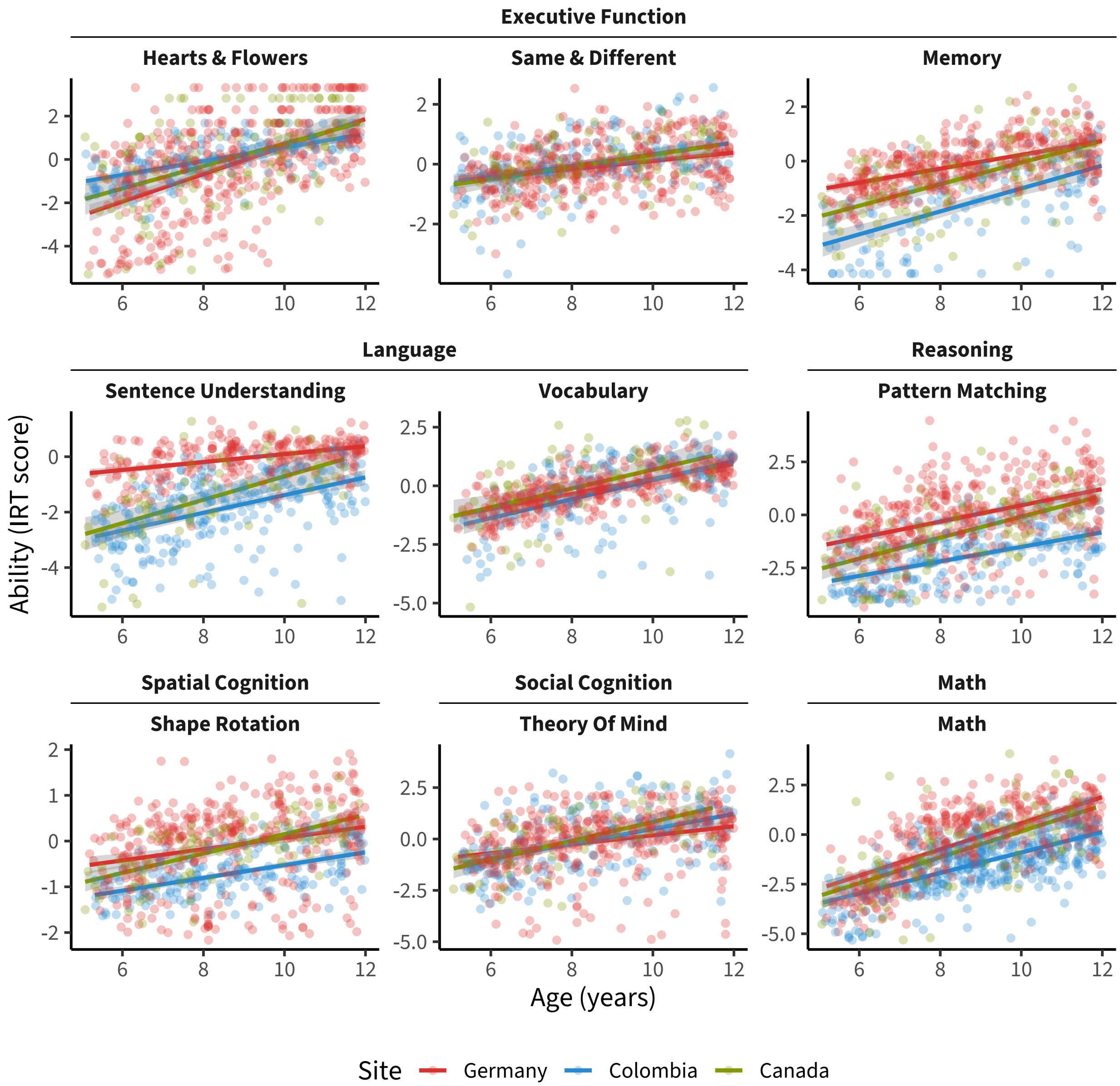
Item-Level Diagnostics

Classical indices included proportion correct (flagged if <0.20 or >0.90) and point-biserial correlations (flagged if <0.20). IRT-based fit statistics included infit and outfit mean squares. DIF Removal and Revision of Problematic Items Ability Estimation CAT-Specific Scoring

Multigroup models and invariance Model selection approach 1PL/2PL and different degrees of invariance Item-level diagnostics Removal of outlier items Tests for item-level DIF in cases where we use multigroup scoring

Psychometric properties of tasks

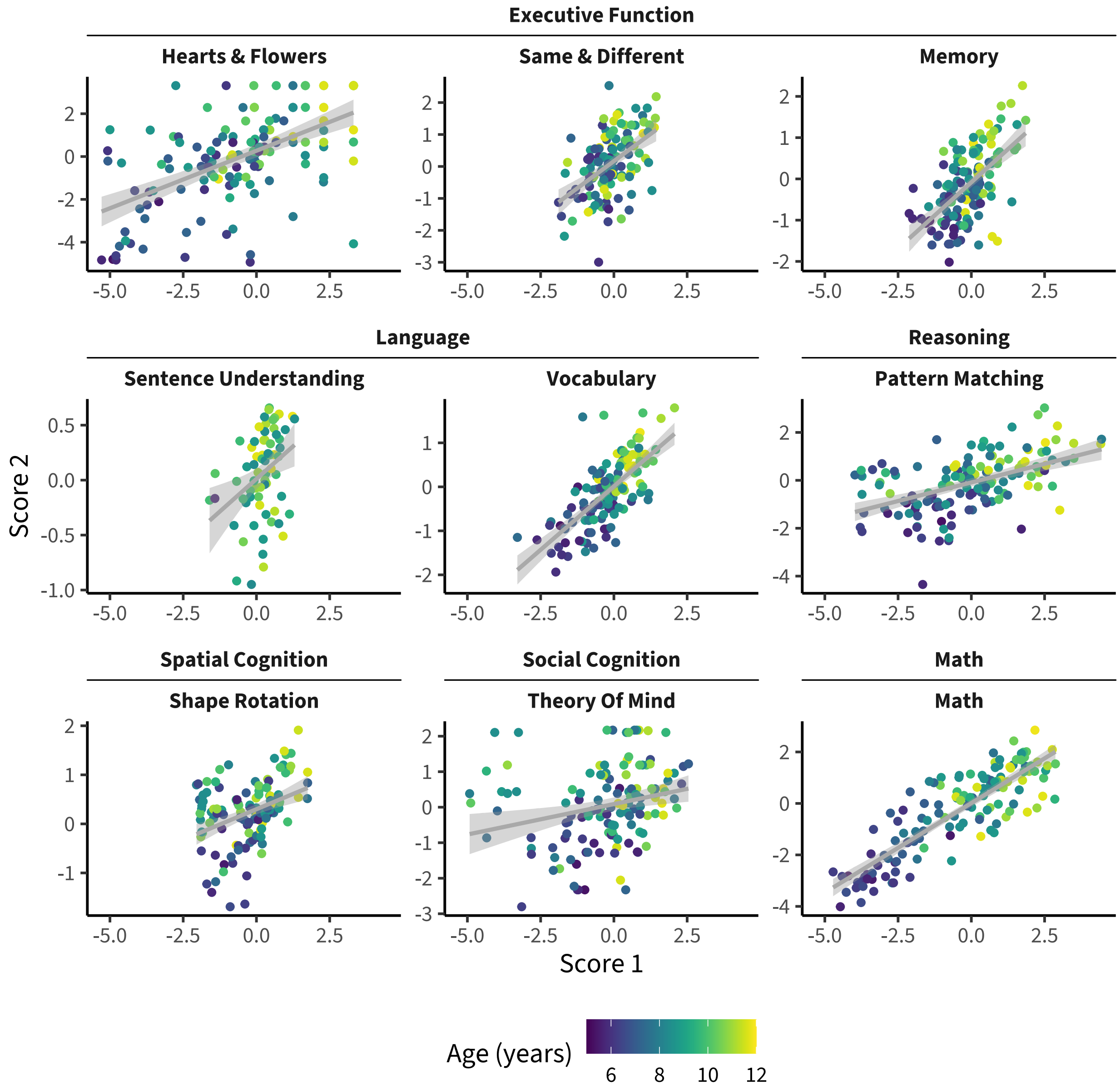
Developmental change



Reliability

Marginal reliability estimate.

Test-retest.



Validity

Construct validity. SEMs within each construct for language, number, EF?

External validity. We next assessed external validity by examining correlations with other measures. In particular, we used …

MEFS and ROAR

Adaptive task construction

Many of the LEVANTE tasks have been adapted and piloted as CATs (Computerized Adaptive Tests). To date, these include the Test For Reception of Grammar (TROG), Vocabulary, Mental Rotation, Matrix Reasoning, Same Difference Selection, and Math. These tasks maintain an ability score, theta, as an estimate of the participant's skill level that is updated at the end of each trial. They then present participants with the item best suited to their estimated ability, which both improves test-taker experience and yields more information on participant skill per item, allowing for a shorter task with fewer items.

The CAT tasks use an adaptive algorithm made available by the jsCat JavaScript library (**ma2025?**), which offers an implementation of an Item Response Theory (IRT) model including up to 4 parameters: discrimination, a value representing the item's informativeness in distinguishing high and low ability test-takers, guessing, the probability of selecting the correct response at random, upper asymptote, the maximum likelihood of a correct answer, and difficulty. The present LEVANTE CAT implementation varies difficulty and guessing for each item while holding both discrimination and upper asymptote constant at 1. Items are selected based on Maximum Fisher Information, and theta is updated according to a maximum likelihood estimator, with limits of -6 and 6.

The LEVANTE CATs are configurable with respect to the initial value of theta and the conditions for ending the task. The starting theta is set at 0 for all CAT tasks currently in use, but can be lowered or raised according to the researcher's prior expectation of participant ability, for example according to age. Current CAT implementations use stopping rules based on either time or number of items. TROG, Mental Rotation, and Vocabulary each have time limits currently set to 4 minutes, with Matrix Reasoning set to 6 minutes to allow for the increased time typically required to complete items in this task. Items in these tasks are presented together in a single block. Same Difference Selection and Math are each divided into three blocks presented sequentially, with per-block stopping based on number of items. These CATs select from the list of items specific to their current block and proceed to the next block once the target number of items is reached, maintaining one overarching ability estimate for the entire task. Same Difference Selection and Math have time limits of 6 minutes and 8 minutes, respectively.

Discussion

- Plans for internationalization
- Plans for downward extension
- Plans for updates

References

Yeatman, J. D., Tang, K. A., Donnelly, P. M., Yablonski, M., Ramamurthy, M., Karipidis, I. I., … Domingue, B. W. (2021). Rapid online assessment of reading ability. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-85907-x

Table 1

The LEVANTE core tasks, presented with their internal label as well as prior labels used in the literature.

Construct	LEVANTE name	Prior names / Source task name	Adaptive?	Reference
Executive Function	Hearts and Flowers	Hearts and Flowers		XYZ
	Memory	Corsi Block Task	X	
	Same and Different	Same Different Selection Task	X	
Language	Vocabulary	Picture Vocabulary	X	
	Sentence Understanding	Test for Reception of Grammar (TROG)	X	
Math	Math	Early Grades Math Assessment (EGMA)	X	
Reasoning	Pattern Matching	Matrix Reasoning	X	
	Shape Rotation	Mental Rotation	X	
Social Cognition	Stories	Theory of Mind	X	

Table 2

Measurement invariance across factor analysis and item response theory.

Goal	LEVANTE name	Factor analysis name	IRT explanation	Interpretation
Exploratory		Multifactor EFA: Groups differ in # of factors and configuration of loadings	Not possible [IRT assumes unidimensionality]	Groups differ in meaning of construct
Measurement Invariance	2PL Configural	Configural: Same # of factors, loadings = 0 for same items	Non-uniform DIF: Groups differ in all item characteristics	Groups express construct in different ways
Measurement Invariance	Rasch Configural		Non-uniform DIF: Groups differ in all item characteristics	Groups express construct in different ways
Measurement Invariance	2PL Metric	Metric (weak): Equal item factor loadings	Uniform DIF: Equal item discriminations (a = a_g)	Groups express construct in similar ways
Measurement Invariance	2PL Scalar	Scalar (strong): Equal item factor loadings & item intercepts	No DIF: Equal item discriminations & item difficulties (a = a_g, d = d_g)	Factor scores are on same scale and can be compared across groups
Measurement Invariance	Rasch Scalar		No DIF: Equal item difficulties (d = d_g)	Factor scores are on same scale and can be compared across groups
Measurement Invariance		Full (strict): Equal item factor loadings, intercepts, residual variances	Not possible [because no residual variances with binary items in IRT]	Factor scores are on same scale and can be compared across groups; Reliabilities also equivalent
Group Differences in Factor Scores		Group Mean Difference: Equal latent variable means	Group Mean Difference: Equal latent variable means	Group A and Group B have same means on the latent construct
Group Differences in Factor Scores		Group Variance Difference: Equal latent variable variances	Group Variance Difference: Equal latent variable variances	Group A and Group B have same variances on the latent construct

Table 3

Task-wise test-retest correlations computed for Germany data.

Task	r	N	Retest interval (months)
Hearts & Flowers	0.56	137	5.08
Sentence Understanding	0.35	68	4.51
Same & Different	0.51	121	4.71
Pattern Matching	0.51	137	4.83
Shape Rotation	0.38	127	4.90
Theory Of Mind	0.25	134	4.82
Math	0.86	134	4.70
Memory	0.59	137	5.03
Vocabulary	0.70	129	4.04