Uncorrected manuscript, accepted at *Annual Review of Developmental Psychology*

Psychometric Rigor for the Real World: Item Banking in Developmental Research

Benjamin W. Domingue*

Abstract

Measurement—a critical component of scientific progress—of attributes in developmental psychology provides a number of challenges. I argue that probabilistic models of item responses resolve many of these challenges and deliver desirable features. Such models allow us to systematically investigate whether measures function equivalently across groups of people, make it possible to deliver consistently meaningful scores from different versions of the measurement instrument (that may be better suited to different ages or developmental periods), and can be used to efficiently maximize the value of time spent with the respondent. To further enhance usability of a measure, its functioning can be codified in the form of an item bank. Item banks, databases containing the information required for reuse of the item, can streamline future measurement tasks given that items can be screened for invariance and selected given the developmental stage of the respondents in question. To illustrate these points, I discuss three recently-developed item banks—the NIH Baby Toolbox, LEVANTE, and the D-score system—meant to manage the kinds of challenges discussed herein.

Significance Statement: Psychological measurement is integral to developmental research. I argue that practices in this space could be improved in concrete ways via an increased emphasis on the understanding of item functioning and codification of this functioning into item banks. I discuss this via a focus on how item banks and the probabilistic models for item responses that they utilize can help facilitate solutions to many challenges associated with psychological measurement (e.g., ensuring that measures are functioning similarly across groups, generating measures that are appropriate for a given respondent). I also discuss three recent examples of item banks—the NIH Baby Toolbox, LEVANTE, and the D-score system—relevant to developmental psychology.

Keywords: psychometrics, item response theory (IRT), measurement invariance, equating, item bank

*Graduate School of Education, Stanford University, Stanford CA, USA; bdomingue@stanford.edu

1 Introduction

Understanding change over time is a central issue in developmental science. Growth in stature is a canonical example. Its measurement with a ruler has several noteworthy properties. A ruler is able to effectively measure the height of a child irrespective of that child's background. An appropriately chosen ruler can rapidly measure changes in stature across the full developmental period (and alternative measurement equipment is available for measuring changes in length in both far larger and smaller increments). The measures generated by a ruler have a consistent meaning that can be readily compared to measures of length generated by different rulers. While the measurement of change in attributes of primary interest in developmental psychology will be more fraught, our intuition of what it means to measure effectively can often be informed by our intuition of how measurement with a ruler should work. One of the key themes here is that we should try to build psychological measures that have the kinds of properties associated with systems for physical measurement (e.g., efficient, scaled for task-at-hand, intervally scaled based on a common unit). This is a demanding task but one that, in my view, is at the heart of building a better developmental science.

As an initial illustration of the challenges we will need to grapple with when measuring psychological attributes, consider measures of executive functioning [Diamond, 2013]. Common tasks used to measure facets of executive functioning (EF) such as a Flanker task may be operationalized on a tablet thus potentially becoming influenced by a child's familiarity with a tablet (to say nothing of the visual requirements of such a task).[1] Children from different backgrounds may have different levels of comfort with a tablet-based task which may then distort our understanding of age-related change in EF across groups. This is but one potential problem. We may also be interested in minimizing the response burden; rapid measurement will require precise targeting given the child's level of development. Moreover, a coherent understanding of 'growth' may require comparisons of measures derived from different tasks onto the same scale (e.g., the tasks required to measure growth in core academic skills vary substantially as children age). This review will attempt to articulate a unifying solution to these problems.

This review considers the key conceptual and operational issues associated with measuring such psychological and behavioral attributes across time and place. After considering these challenges, I argue that there is one unifying approach to constructing measurement systems that could help propel psychological measurement forward. In many respects, the specifics of this call are not new—aspects of it date back several decades [Embretson, 1996]—but it is timely given the increasing amount of psychological measurement done via digital device and the interest in increased understanding of development

[1] Indeed, alternative measures of EF that do not require tablets have been developed for scenarios where tablet-based assessment may be inappropriate or infeasible [Molfese et al., 2010].

in non-WEIRD settings [Henrich et al., 2010].

I consider a probabilistic understanding of item functioning as central to building systems that have the desired features. Having probabilistic models for item responses has many advantages. First and foremost, such models can be falsified; we can interrogate how such models are working in a given scenario. Second, as we shall see, such models can be used to diagnose violations of measurement invariance and to resolve other practical testing problems. Finally, when such models are judged to function effectively, they allow for concise description of a measure's functioning that can be saved for subsequent reuse. Specifically, this kind of description when codified into an 'item bank' can help practitioners deal with the challenges—e.g., need for rapid assessment, questions about portability across time and place—faced in practice. Rather than having numerous different measures used for different ages and contexts, an item bank can be a unifying force. An item bank containing a reasonably large number of items with established and documented measurement properties can be used to compose tailored assessments for different populations, age groups, or waves in a way that maximizes information, enhances inference, avoids bias, and ensures comparability across people, samples, and time. Alongside a conceptual discussion of item banking as a sound platform for resolving many measurement challenges, I will also describe three recent examples of measurement systems focused on children that have such solution-oriented features [Gershon et al., 2024, Frank et al., 2025, Weber et al., 2019] as instantiations of how item-focused measurement systems can be developed and used.

Item bank: A database containing information about both the item, its statistical properties, and other information needed to ensure appropriate reuse of the item.

2 Conceptual Considerations

While my main goal is to emphasize solutions to some frequently-encountered (if my experience consulting on measurement challenges in education and psychology is to be trusted) problems in psychological measurement, I begin with critical historical context for these solutions. Psychological measurement is a difficult endeavor and a better understanding of some of the underlying philosophical disputes will help practitioners understand the limitations of psychometric techniques. The issues that I discuss here are largely conceptual but are relevant for my goals of offering real world guidance given that they clarify the limitations we should anticipate of our measurement systems. I will tackle three questions—what does it mean to measure? what kind of scale properties should we anticipate? how do we think about validity?—that have been of historical importance and remain of contemporary relevance.

2.1 The meaning of measurement

The classic definition is that measurement is the "estimation of the ratio between a magnitude of a quantitative attribute and a unit" [Michell, 2014] (the

modern definition of measurement is more complex [Mari, 2013]). However, whether psychology should aim to comply with this definition has been contested over the years [Stevens, 1946, Michell, 1997]. My aim is to foreground the issues raised in these debates that lead to a need for humility when considering psychological measures (deeper discussions of these debate are available elsewhere [Briggs, 2021, Michell, 1999]). The core challenge is that the majority of constructs of interest in developmental psychology are not things with which we can interact directly but rather 'latent' in the full complexity captured by Bollen [2002]. An early expression of concern from 1940 regarding whether 'measurement' was possible in psychology comes from the Final Report of the Ferguson Committee [Ferguson et al., 1940] which objected to the notion of measurement in psychology given that psychological constructs cannot be concatenated (i.e., I can put two one foot rulers together to make a two foot ruler but I cannot in any meaningful way combine executive functioning across tasks or people).

Reactions to this report varied. One was to propose a more flexible definition of measurement (i.e., Stevens' levels of measurement) premised on measurement as "the assignment of numerals to objects or events according to a rule" [Stevens, 1946]. Separately, developments in the field of conjoint measurement [Krantz et al., 2006] demonstrated that concatenation is not essential and that, under certain assumptions, scales with desirable properties may still be possible even without concatenation. Let us consider each of these reactions.

2.1.1 Measurement versus classification

Steven's framework [Stevens, 1946] emphasized several different levels of measurement. For our purposes, we shall consider nominal and ordinal scales for measurement as 'classification'.[2] One theme here is that, consistent with measurement of length, we should strive for measurement in the classical sense while also acknowledging that this is no small feat with psychological attributes. Given this goal, my reaction to classification-based is typically a query of "can we do better?". Theory often anticipates that the latent classes of interest may be partitions of a unidimensional or multidimensional scale (thus suggesting we do not require the class-based approach) and the statistical tools we use to make inferences about classes frequently confuse continuous variation for class-based variation (Bauer and Curran [2003] provide an example from the literature on growth mixture models). We will ultimately have both improved theories and applied results if we attempt to map out and

[2]In not referring to these as 'measurement', I am going counter the guidance offered elsewhere [Torres Irribarra, 2021]. My view is that the word measurement has an implied meaning that we should acknowledge. If I were to identify the species of various insects outside my home, saying "I am going to measure the bugs." would almost certainly lead someone to misunderstand what I was going to do. Classification would match peoples' expectation of this activity.

measure the relevant continuous scales rather than immediately resorting to partitions of them.

Even on a more practical footing, there may be reasons to move beyond class-based approaches. As one example from my own work [Ma et al., 2025a], we considered the relative performance of diagnostic classification models—a class-based item-response theory approach [von Davier and Lee, 2019]—when compared to conventional approaches relying on an assumption of continuous abilities (similar to Eqn 1 below). We first show, in simulation studies, that these class-based approaches offer superior predictions as compared to continuous alternatives when they are used to generate data. This evidence, when combined with the observation that such approaches can flexibly fit even arbitrary data [Bonifay and Cai, 2017], suggests that class-based approaches might be expected to outperform continuous approaches in empirical settings. This is not at all what we observe. Rather, continuous approaches produce superior predictions relative to class-based approaches (in data specifically chosen given prior usage with class-based analyses) and are far less susceptible to overfitting. Such evidence should make us reluctant to work with such class-based approaches absent a compelling rationale.

As an alternative to direct modeling of classes, I would encourage attempts to first build scales for the relevant attributes and to then interrogate the scales. If there are behavioral profiles that do not map in intuitive ways onto the scale, it might be that class-based approaches would be superior. In other cases, classification based on a cutpoint, as in 'standard setting' [Cizek and Sternberg, 2001, Robinson, 2011], may be a useful alternative for attempting to identify regions of the underlying continuum which represent qualitatively different types of performance. This encouragement is not meant to suggest that there are no valid uses for classification (see relevant discussion in Torres Irribarra [2021] and the example in Borsboom et al. [2004]). Rather, we should be striving to derive more functional approaches to measurement when we can. This means working to build scales that go beyond classification or, when that is not possible, to better understand the qualitative structure of the relevant attribute that necessitates classification. Even if research does utilize class-based approaches, one must be clear-eyed about this being a potential limit of our provisional understanding of the attribute we are studying rather than a ground truth (i.e., we used to classify hardness via Moh's scale but it can be measured intervally using more mature techniques [Ghorbani et al., 2022]).

2.1.2 Interval scales

Interval scales have a desirable property. A ruler with markings that are all equivalently spaced and representative of some unit (e.g., an inch) is intervally scaled whereas a ruler with differently sized markings is not (see Figure 1 in Briggs [2013] for additional discussion on this point). Note that the unit emphasized here is the same unit emphasized in the classical definition of measurement in Section 2.1. The length scale is defined with respect to a

standard unit (e.g., an inch in the British imperial system or the meter in the metric system) and all markings need to be faithful replications of this unit. No matter the lengths of the objects in comparison, we can always evaluate their differences by comparison to the base unit. If we consider, for example, a simple sum score [Sijtsma et al., 2024], differences do not have an interpretation in terms of reference to some standard unit. Indeed, most psychological scales do not even define a reference unit, the Lexile [Stenner et al., 2006] and D-score [Weber et al., 2019] scales being notable exceptions.

Interval scale: A scale where differences are consistently meaningful in terms of the underlying unit.

When might interval scales be possible with psychological attributes? Developments in 'additive conjoint measurement' [Luce and Tukey, 1964] described a specific set of conditions sufficient for ensuring an interval scaling of an attribute. Speaking roughly, these conditions work to ensure independent variation in the items used to measure and the things being measured influence a response. The task then becomes to test consistency with these conditions. Direct tests exist [Karabatsos, 2001, 2018, Domingue, 2014, Davis-Stober, 2009] but it is also known that the Rasch model which we discuss below meets these conditions [Perline et al., 1979]. Consistency with either the Rasch model or additive conjoint measurement requires assuring that measurement data have certain types of structure. It can be difficult to refine measurement systems such that they produce data that meet the relevant criteria yet I encourage psychologists to embrace this challenge given that attempting to make such refinements will lead to improved understanding of a given attribute. Failing to develop interval measures is also likely to mislead: even if there is little rationale for why a given scale might be interval, symbolic placeholders for scale levels are often transmogrified into cardinal numbers when people start using them. Put differently: I can assure you that people will calculate differences between those symbolic placeholders. As in other disciplines [Sherry, 2011], efforts in the direction of constructing interval scales will be integral to improved psychological science. Interval measurement is consistent with colloquial usage of the term measurement, allows for analysis of differences in scale values, and permits straightforward analysis via the common statistical tools used by psychologists; it is the target we should be aiming for.

Rasch model: A family of item response models that meet stringent criteria [Fischer and Molenaar, 1995]. An example is in Eqn 1.

2.2 Validity

It is essential for users of a given measure to argue that it is 'valid' but this necessity has led to a shifting debate about the meaning of validity. Let us begin with the stronger version of validity codified by various professional organizations which emphasizes use: "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." [Joint Committee on the Standards for Educational and Psychological Testing, 2014, p. 11]. Interpretations and uses of test scores, which are external to the measure, have starring roles here. This emphasis is especially salient when such scores are used for high-stakes decisions. For example, if a decision about whether a child should receive some kind of intervention based

on the outcome of some measurement process, evidence about the efficacy of the intervention for *this* child as compared to other children will be essential [Robinson, 2011]. If all children benefit from the intervention, we may not need the measure. If, on the other hand, children at this level of the outcome tend to benefit from the intervention but others do not, the inference is on much firmer ground.

Validity: A topic related to the use and interpretation of psychological measures but one whose definition is contested.

While there is clearly a need to analyze whether uses of a given measure are appropriate, emphasizing external criterion such as uses simultaneously undermines consideration of internal criteria and asks developers of measures to account for use cases that they may not have control. This has prompted articulation of an alternative view that focuses on internal features: "A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes" [Borsboom et al., 2004]. The connection between the attribute and the measure is central. For lower-stakes work, this internal notion seems like the more obvious focus given that uses are de-emphasized. Further, when we think about measurement in the physical world, we think about the internal qualities of that process not how it is used; understanding whether a ruler is a valid measure of length requires no information about subsequent uses. This definition does not immediately offer guidance related to the fact that variation in non-focal attributes may also cause variation in the measurement outcome. The authors consider this point (see p. 1070) but I will emphasize here that we should generally standardize assessment conditions, to the extent possible, to minimize the undue influence of non-focal attributes.

Validity is a thorny subject. Issues of 'use' will largely be beyond the scope of this discussion; I will focus on the inner-workings of a psychological measure. Much of that discussion focuses on obtaining highly 'reliable' measures in the sense of minimal measurement error. Given my goal of offering pragmatic guidance, I want to now turn to a few key criteria for successful measurement that, if met, will help to support even the broader notion of validity.

Reliability: A quantity indexing the precision of a psychological measure.

2.3 Key desiderata for psychological measures

We carry substantial intuition regarding what we desire from measurement systems based on (even limited) experience with measurement of properties of physical objects (e.g., length, mass). These intuitions are useful in both identifying desirable features of measurement systems in psychology while also making clear the limitations of existing technology in this realm. What are these key desiderata? I emphasize three that will be relevant:

KEY DESIDERATA

1. Measurements should be insensitive to differences in nonfocal attributes.

2. Measurements should be comparable across measures taken from alternative approaches.

3. Measurement systems should provide precise information about the attribute as rapidly as possible.

While no system of psychological measurement can be expected to have perfectly achieved these goals, my thesis is that an increased focus on the probabilistic understanding of the items which constitute many measurement systems will result in progress towards them. To demonstrate why this is so, I begin with a discussion of how probabilistic models can be used to study data derived from psychological measures.

3 Items and item banks as the foundation of an operational measurement program

3.1 A probabilistic model for item responses

For the purposes of connecting the various issues discussed to measurement practice, it will help to be specific. We consider a single model—the Rasch model [Rasch, 1993, Fischer and Molenaar, 1995] which is the simplest item response theory (IRT) model—before broadening this perspective in Section 4.4; in particular, we start with dichotomously coded responses before expanding to polytomous responses. If person i responds correctly (1) or incorrectly (0) to item j, suppose that

$$\Pr(x_{ij} = 1|\theta_i) = \frac{1}{1 + \exp(-(\theta_i - b_j))} \tag{1}$$

where θ_i is a scalar representing person i's location on some latent scale and b_j is the difficulty of item j. The difference between θ_i and b_j controls the probability of a correct response (see left panel in Figure 1). When $\theta_i - b_j >> 0$, the response will almost certainly be correct whereas when $\theta_i - b_j << 0$ the response will almost certainly be incorrect. In this model, the only difference between respondents is the θ_i quantity while the only difference between items is their difficulty b_j. The model offered in the left hand of Figure 1 will be the key to building item banks which solve many measurement issues.

Eqn 1 is built on a few key assumptions that merit comment.

- Responses are assumed to be independent in that the response x_{ij} does not depend on responses to other items j conditional on θ_i, b_j; this is a standard assumption with many latent variable models [Bollen, 2002]. Techniques exist for assessing this assumption [Edwards et al., 2018] and

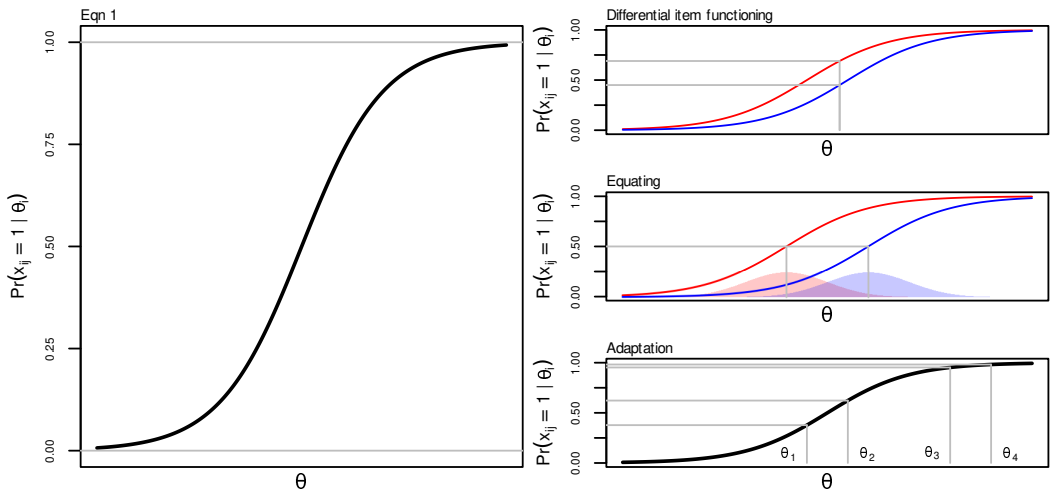

Figure 1: Left: Illustration of expected response behavior ($\Pr(x_{ij} = 1|\theta_i)$) as a function of θ_i for a single item (i.e., a fixed value of b_j). The probability of a $x_{ij} = 1$ response increases as a function of θ_i. Right: (Top) Differential item functioning: The same item is uniformly harder for the respondents in group represented by blue IRF as compared to the respondents in the red IRF. (Middle) Equating: An easy item (shown in red) is given to respondents with lower θ values; this results in an average response of 0.5. A more difficult item is given to a group with larger θ which also results in an average response of 0.5. With just the average we cannot make any claims about the relative difficulties of the items or abilities of the groups (i.e., the horizontal gray lines converge). IRT scales utilize the notion of parameter invariance to partition between group ability and item difficulty. (Bottom): Adaptation: The values are chosen such that the differences $\theta_2 - \theta_1$ and $\theta_4 - \theta_3$ are equal. This item does little to differentiate between θ_3 and θ_4 respondents given that their expected responses are quite similar. In contrast, this item would be more appropriate for respondents in the region around θ_1 and θ_2.

there are scenarios where it needs to be relaxed (i.e., testlets; Wainer et al. [2007]). Reading comprehension measures, for example, may have a person first read relatively long passages and then respond to several items thus embedding potential dependencies amongst these responses.

- Eqn 1 assumes that the person-level attribute θ_i is unidimensional; we return to this assumption in Section 4.4.

- Item functioning only depends on a single parameter b_j. There are a variety of more flexible variants of Eqn 1; conventional alternatives would be to model responses with two or three item parameters [Lord and Novick, 1968] but even more sophisticated variants exist [Shim et al., 2025, Van Schuur, 2003]. Much of my recent work has focused on comparing the performance of such models [Stenhaug and Domingue, 2022,

Domingue et al., 2024]. One conclusion of this work is that the differences between these models can be fairly muted and we may not always need to fret about selecting the optimal model.

- Eqn 1 assumes that the probability of a correct response increases monotonically as a function of θ. This is desirable for some but not all items.[3] Models wherein the probability will decrease as the distance between θ_i and the relevant item parameter (i.e., an alternative to b_j) increases are available for such 'unfolding' style response processes [Andrich, 1996].

Before moving forward I want to emphasize an important conceptual question: how much flexibility do we want in our measurement systems? The idea of 'specific objectivity' is an important test. It asks whether comparisons of objects generalize beyond the specific instruments used to compare them. For example, inferences about who is tallest amongst a group of people shouldn't hinge on which ruler is used to make the inference. A violation of specific objectivity raises question about whether we are 'measuring' in the conventional sense. The Rasch model, if appropriate, ensures specific objectivity holds which means that respondent orderings are preserved across any set of items used to compare them. Specific objectivity does not necessarily hold when more complex functional forms must be used in place of Eqn 1; whether such flexibility in modeling is desirable has long been a source of contentious debate [Wright, 1997] and is closely linked to the discussion in Section 2.1.2.

Specific objectivity: Comparisons of objects should generalize beyond the specific instrument(s) used to compare them [Rasch, 1993, Fischer and Molenaar, 1995].

After collection of responses by respondents to items, estimates of θ_i and b_j can be derived. Frequentist approaches typically separate estimation of item and person parameters. Item parameters (e.g., b_j) are first estimated using a variety of techniques [Baker and Kim, 2004] with expectation-maximization approaches [Bock and Aitkin, 1981] being a suitable default. Once item difficulties are estimated, person parameters (e.g., θ_i) can then be estimated in a relatively straightforward way using maximum likelihood or related techniques. Bayesian approaches are also available [Fox, 2010]. Note that missing responses can be accommodated given that Eqn 1 focuses on modeling of a response from person i to item j (e.g., we can still derive estimates for a person i if they do not take all items). Given that Eqn 1 is invariant under a translation of θ_i and b_j, we need to fix the scale by making some assumption. This is often done by fixing the distribution of θ_i to have a mean of zero and standard deviation of 1.

Let us now suppose that assumptions of Eqn 1 are met and that data has been collected in a first group G_1 which allows for estimates of θ_i and b_j. At this stage, researchers can interrogate the degree to which Eqn 1 (or some alternative) actually describes empirical observations. If it does not, this may be a signal that the measure needs to be improved via an iterative process

[3]Items in affective measures such as "A dinner with a few friends makes for a perfect Friday night." where respondents who want to be alone or those who want to party will both reject such a statement.

(e.g., items may need to be pruned if the Rasch model is to be used [Wu and Adams, 2013]). If it does, then I think—agreeing with others [Elson et al., 2023] that increased reuse of psychological measures is an important scientific aim—**researchers can greatly increase the usability of their work by codifying their findings in an item bank**. In Section 3.2, I further discuss item banking (i.e., the codification of what is known from study of G_1) before then discussing their use to solve common measurement problems.

3.2 The joys of item banking

An item bank is a database containing both information about the item as well as information about its parameters (e.g., estimates of b_j from Eqn 1). I will emphasize the affordances of an item bank for a measurement program and leave additional technical considerations, especially germane for higher-stakes or larger-scale use, for others [Vale, 2006]. Item banks can be tightly-held internal resources. For example, when items from the banks used in high-stakes standardized testing situations like the SAT are exposed, this can lead to real problems. In other scenarios, the item banks can be shared publicly (as in Section 5). Such resharing will be appropriate for many low-stakes measures used in developmental psychology and integral to subsequent research and practice.

How are item banks developed? Typical research projects in psychology collect data and then use the same data to study both the measure and the respondents. To understand an item bank, it may be helpful to distinguish between data collection done for the *purpose of understanding items* versus the *purpose of understanding respondents*. With a calibrated item bank, one is treating the items as fully understood and subsequent data collection will focus purely on delivering information about respondents. This, of course, hinges on data collection first being undertaken for the purpose of understanding the items (i.e., the work done with group G_1 above). This initial data collection allows for identification of item parameters. Subsequent work then uses this fixed item functioning to understanding respondents (i.e., estimation of θ_i).

Item banks are a valuable organizing structure in ongoing measurement systems. As with any structure, they require maintenance. Some items may need to be edited or removed over time. Editing items, or even introducing new items to the bank, will require approaches for getting updated parameter estimates that can be stored in the item bank Ren et al. [2017]. In high-stakes settings, knowledge of the items in the bank can lead to inflated scores. Changes to item parameters over time are known as 'drift' [Bock et al., 1988] and need to be managed. A related issue is management of 'exposure control' so that certain items are not overused in adaptive testing scenarios [Van der Linden and Choi, 2020]. Finally, when collecting data to establish an item bank, care must be taken so that practice or order effects don't influence parameter estimates. This can typically be done by giving items in varying order to respondents.

Suggestions for reporting on new measures in developmental psychology

Given the increased importance of open science practices in psychology, the following approaches to reporting on novel measures in developmental psychology may help increase reuse of such measures. Researchers should:

1. Offer a complete description of each and every item.
2. Report on item functioning using both descriptive statistics but also parameter estimates from statistical models (e.g., using something akin to Eqn 1).
3. If possible, release response-level data. This will allow subsequent researchers to investigate measurement invariance and also to fit other measurement models of interest.

4 Item-based solutions to common measurement challenges

4.1 Measurement invariance

Measurement invariance [Millsap, 2012] implies that Eqn 1 governs response behavior irrespective of whether respondent i is in G_1 or G_2. Merely building an item bank won't ensure that measures are invariant but, rather, an item bank makes assessment of measurement invariance an empirical question that can be scrutinized. Violations of measurement invariance are a fundamental threat to the validity of a given measure if it is to be applied broadly [Joint Committee on the Standards for Educational and Psychological Testing, 2014] as such violations can be closely tied to the issue of fairness [Camilli, 2006]. In particular, construct-irrelevant variance in measurement outcomes that are associated with some other feature of the respondent are particularly concerning. Analysis of invariance is common practice with standarized tests in education but perhaps less so in psychology more generally [Maassen et al., 2023]. Below, I discuss several issues germane to such investigations.

Measurement invariance: Consistent functioning of a measurement instrument across contexts.

There are a number of ways in which measurement invariance may be violated. We focus on a common and particularly concerning violation of measurement invariance: the difficulty b_j does not accurately describe the difficulty in both groups of respondents (i.e., uniform differential item functioning). A visual depiction is shown in the top panel on right side of Figure 1; the two curves represent differential functioning of *the same item* for two groups. Respondents from the blue group have a uniformly more challenging time on this item conditional on ability (i.e., for a fixed point on the x-axis, the blue

curve is always lower). This is a violation of the assumption that θ_i is the only person-level attribute that is impacting performance on x_{ij} as, clearly, group membership also matters.

Suppose we are assembling a set of words to measure a child's capacity for language (by seeing whether they can produce them when given appropriate stimuli). Consider the word 'horse'. If G_1 is US children, Wordbank [Frank et al., 2017] norms suggest that roughly 70% of children will know this word at 24 months (this percentage is related to the value of b_j). However, there is substantial global variation in the proportion of 24 month olds who know this word: just over 50% of the children in Russia or Argentina would be expected to know 'horse' as compared to 90% of the children in Beijing or the Netherlands. 'Horse' is just one word out of many that we could use to assess a child's knowledge of language and the degree to which children know it will vary as a function of (presumably) their cultural context and the nature of that word in their language (e.g., is it especially long or hard to pronounce? are horses common in their culture?).

Variation in the relative difficulty of a word across groups will lead to between-group variation in b_j that is a violation of measurement invariance. The violation of invariance in this example hinges on the assumption that 'horse' is an idiosyncratic indicator of the attribute that we are aiming to measure rather than its essential feature. As a contrast, if we were asking about the proportion of children aged 2y greater than 80cm in stature, between-country variation in this proportion is not indicative of a problem with how length is measured given that there is no ambiguity about the essence of the thing we are aiming to measure.[4] Note that I have described only a relatively straightforward type of invariance focusing on item difficulty but other violations of invariance also exist.[5]

The above tacitly assumes that variation in knowledge of 'horse' is not indicative of broader language-development differences between cultures and is instead word-specific (i.e., I am comfortable assuming that children at 2y have similar language abilities in most places). We will often be explicitly interested in item-level differences in functioning with groups for whom there are overall performance differentials and such an assumption would not hold. How do we detect measurement invariance when we cannot rely on this assumption? The key is to identify something that is invariant between the groups. If, for example, the groups are formed by random assignment we can assume that the distribution of respondent abilities are equivalent. This may be viable in studies of mode effects where random assignment to assessment modality is possible [Domingue et al., 2023] but is impractical for groupings based on

[4]Although, even with anthropometric measures, cross-cultural comparisons can be fraught [Borghi and Sachdev, 2024].

[5]In applications of IRT approaches, uniform and nonuniform DIF are considered when dealing with dichotomous items [Camilli, 2006] and matters are more complex with polytomous items. In SEM analyses, there is a separate typology of violations in widespread use [Putnick and Bornstein, 2016].

(nonrandom) features of a person. More typically, some set of items must be identified that function equivalently. A common approach is, when studying one item, to assume invariance amongst all others. This is a non-trivial assumption and recent work has advanced novel approaches for selecting more refined subsets [Halpin, 2024, Belzak and Bauer, 2020]. Having first identified an invariant feature, statistical approaches are then used to examine whether group membership is leading to a variation in measure functioning (for us, Eqn 1). A variety of approaches [Camilli, 2006, Section 7] and software tools [Magis et al., 2010] exist for conducting such analyses.

When measurement invariance is found not to hold, caution is merited. If the groups G_1 and G_2 are, for example, based on respondent age, then evidence for noninvariance may be evidence that the nature of the construct is changing as respondents age. One can also observe shifts in the nature of the construct for those who receive treatment in a randomized trial [Gilbert et al., 2025]. In educational measurement, it is common to remove items that exhibit bias (in the sense of DIF). As one example of a common problem related to a failure of invariance, complex language used in math problems on academic assessments can lead to differential item functioning for respondents assessed in their non-native language [Buono and Jang, 2021, Solano-Flores et al., 2013]. Whether such items should be included then becomes a question related to the desired inference: if the assessment is meant to purely assess math ability and the word problems are nonessential, such items should likely be removed whereas if interest is in the performance of mathematical computations in context the question is more difficult.

Researchers should resist credulity and, when we are measuring psychological constructs across settings, should have a prior that invariance will not hold as settings become more distinct. Observed violations may be useful in helping to refine our understanding of the construct. But, they also offer ripe opportunity for exercising humility. I close this section with a maxim I have used while teaching and that may resonate with others: psychological measures are a useful tool for understanding how people who start in similar situations end up different. They are a terrible tool for understanding why people who start in different situations don't end up the same.

4.2 Equating

We frequently cannot give respondents the same items and yet we may want their overall performance—in the form of θ_i estimates—to be comparable. This can be a challenge. Suppose all we know is the the average response to an item (see middle right of Figure 1). The red item is easier but when given to lower-ability respondents produces the same overall level of accuracy as the more difficult blue item given to higher-ability respondents. Thus, the average alone cannot help us distinguish between item difficulty and respondent ability. However, if we are using items drawn from a bank, this problem is effectively solved as the item bank defines the ability scale via the item parameters.

Obtaining comparable ability estimates across respondents irrespective of the items is a benefit of using an item bank.

However, there are other quite common scenarios wherein respondents take different sets of items prior to their calibration (i.e., an item bank is not yet established) with the aim of putting scores on a comparable scale. This can be done using 'equating' techniques which rely on probabilistic item response models. Equating is fundamentally a design problem; once the appropriate design is identified and used for data collection, the subsequent analytic techniques are typically straightforward. Given this, I will briefly discuss some general principles for equating and a straightforward equating design (a broad array of designs exists, see Kolen and Brennan [2013]).

Equating: The act of mapping disparate scores onto a common scale.

Supposing that a set of items—such sets of items are typically described as 'forms'—has been given to students in group G_1, the question at-hand pertains to how responses to a different form can be collected from students in G_2 such that estimates for respondents from each group are comparable.[6] We shall suppose that groups G_1 and G_2 are nonequivalent (the key ideas are also then readily applied to the simpler case of randomly equivalent groups and, in that simplified case, ideas related to 'observed score equating' [Livingston, 2014] apply). Speaking generally, these techniques will be safer when the groups G_1 and G_2 are 'nearly equivalent'. A common application in education pertains to linking end-of-year standardized assessment scores across grade levels. For many reasons (e.g., some items may not be developmentally appropriate, there are different content standards at each grade-level), we desire to give students of different ages somewhat different items. Usage of appropriate equating techniques—typically referred to as 'vertical equating' in this scenario—to make scores between grade 3 and 4 students is one thing, trying to make scores comparable between grade 3 and 10 students another.

Let's consider the specifics of what an equating provides and some general principles for when it can be done. Suppose that a person from G_1 responded to a specific form and got a score of y_1 (i.e., their estimate of θ_i which may have been linearly transformed). An equating is a map that tells us what the equivalent score from the G_2 form would have been; let us call this map f and suppose that $f(y_1) = y_2$. There are many potential ways of producing such a map. Regression, for example, could be used to construct such a map with suitable data. However, equating maps are required to be symmetric (to avoid regression to the mean) such that the map of scores from G_2 to G_1 should be the inverse of f. There are other requirements on both the forms and the groups G_1 and G_2 that may be relevant but one additional requirement merits emphasis: it should be a matter of indifference to the examinee which form they encounter (see Section 13.2 in Lord [1980]). There is intuitive appeal in this rule. For example, if measurement error for scores in G_1 are twice as big

[6] There are several terms (e.g., equating, linking) used for the general activity of putting scores onto the same scale. These terms are meant to cover somewhat different scenarios [Kolen and Brennan, 2013].

as for scores in G_2, then this principle is violated.

The key idea is that there should be some overlap between the items that the two groups receive. Let us suppose that the overlapping items are J_*; these are known as the 'anchor' or 'common' items (the approach is often described as 'non-equivalent groups with anchor test' or NEAT). Everything will hinge on the items in J_*. Roughly speaking, we want items that span as much of the construct as is appropriate given the nature of G_1 and G_2. If a certain developmental task is emphasized in the curriculum received for those in G_2 but not in G_1, then items focused on such tasks will likely not be considered for inclusion in J_* given that they are inappropriate for G_1 students but this is a limitation of the equating that we must then consider. A second consideration pertains to the number of common items. Conventional guidance [Budescu, 1985] suggests that 20% of a form be anchor items (up to 20 items in the case of longer forms) but simulation studies tuned to specific conditions can also help inform design (less overlap may be appropriate for lower-stakes use).

Having collected data from respondents in G_1 to their items and respondents in G_2 to theirs, what next? We will leverage the fact that response-level missingness can be ignored in estimation of IRT parameters. If response-level data from both groups are available, they can be combined into a single response matrix (that will have some structural missingness) and estimation can proceed using multiple group 'concurrent calibration' [Bock and Zimowski, 1997]. This approach is straightforward given that it is effectively just standard estimation of the parameters for an item response model. When response-level data is available for equating, it is the approach I would recommend (and its robustness can be assessed via application of parameter-based equating described below if needed). Indeed, concurrent calibration can be used in cases wherein data are collected such that each respondent sees a unique list of items—suppose that each of 1000 respondents gets a random list of 20 items chosen from a bank of 50 items—so long as there is overlap in these lists across respondents. This approach does depend on having response-level data available. If response-level data is not available, one can produce equatings using item parameter estimates via alternative techniques [Stocking and Lord, 1983].

4.3 Adaptive Testing

One of the most appealing features of an item bank is that it can reduce the response burden by 'adapting' the sequence of items to the respondent so that information is collected efficiently (see bottom right of Figure 1). For respondents with abilities θ_1 and θ_2, we would be interested in their responses to this item given that their ability differences manifest as more distinctly different expected responses (on the y-axis) given the difficulty of this item. If given to respondents with abilities θ_3 and θ_4, this item will not be able to differentiate between them as the expected responses are nearly identical. Another item—specifically an item with greater difficulty—would be more suitable for comparison of these respondents.

Given that time with respondents is often the most precious resource, efficiency is important. Adaptation—typically described as 'computer adaptive tests' or CATs [Chang, 2015]—works by identifying the item that will be most valuable for improving our understanding of respondent ability given a provisional ability estimate based on their previous responses. Harkening back to our earlier discussion of the word 'horse', by 48 months we can anticipate nearly all children will know this word and thus observing that an individual child does will not be very informative about their ability. We would be making more efficient use of limited time with respondents of this age by using a word that is more carefully tailored to distinguishing between such respondents. While there are a variety of specific approaches for deciding amongst the items in the bank, the basic idea is to choose an item whose difficulty b_j is relatively close to the current estimate of ability $\widehat{\theta}_i^*$ (where the $*$ emphasizes that this is a provisional estimate). Practically speaking, the differences between the various approaches to item selection are marginal compared to moving from fixed to adaptive testing approaches in the first place. CATs rely on computers but, even in simpler settings, the relevant ideas can be used to tailor fixed 'short' forms if necessary.

What kinds of efficiency gains can we expect? As a specific example, we used adaptive approaches to streamline delivery of lexical decision tasks in the context of a reading screener [Ma et al., 2025b]. With an adaptive assessment, we were able to achieve a reliability of 0.9 after administering only 75 items versus 125 required for randomly delivered items. The increased usage of digital devices for measure administration and the availability of software [Chalmers, 2016, Magis and Raîche, 2012] through which CAT approaches can be deployed make incorporation of adaptive testing ideas into psychological measurement an appealing proposition!

4.4 Additional Caveats and Considerations

The one parameter IRT (Rasch) model described by Eqn 1 is used for purposes of clarity; I now discuss alternatives to Eqn 1 and how their use may require modification of the above commentary. Suppose that x_{ij} is scored in more than two categories. When x_{ij} is polytomous, there are a range of appropriate IRT models [Tutz, 2020]. When responses are scored in K categories, such approaches focus on modeling $\Pr(x_{ij} = k|\theta_i)$ for $k \in \{0, ..., K-1\}$. Approaches for studying measurement invariance and equating are possible but are more complex given that such models typically utilize more parameters [Nering and Ostini, 2011]. Despite these complexities, the key ideas emphasized here persist whether x_{ij} is dichotomous or polytomous given that they are premised on similar usage of probabilistic models.

A separate consideration pertains to whether the individual-level construct in question is unidimensional or multidimensional. Multidimensional approaches are available [Reckase, 2006] but I argue that, to the extent possible, measurement should be done with unidimensional constructs. My perspective is that

more refined systems of measurement focus on one thing at a time. When one goes to the doctor, they measure height and weight rather than the more amorphous 'size' used for production of cheap clothes (indeed, even more bespoke apparel options typically rely on a a range of unidimensional measures). While I think the intuition behind this is straightforward I realize that this is advice counter to the tradition in psychology related to the use of factor analysis tools for analysis of multidimensional data. In my idealized view, factor analysis tools—along with other approaches to understanding the dimensionality of a set of item responses [Stout, 1987, Buja and Eyuboglu, 1992]—could be used to parse attributes into their constituent parts which can then be measured separately. Developmental science can then focus on measuring unidimensional attributes and using the resulting measures to understand their interrelationships.[7] For an example along these lines, consider Figure 5 in Kachergis et al. [2025].

In the factor analysis or structural equation modeling traditions (I'll refer to these collectively as factor analysis), the goal is modeling the covariance matrix (consisting of elements $\text{cov}(x_{\cdot j}, x_{\cdot j'})$) rather than the item responses x_{ij}. While there are mathematical connections between these approaches [Takane and De Leeuw, 1987], there are also some distinctions. The factor analytic tradition tends to emphasize multidimensional analysis. As noted, I prefer attempts to separately measure and model each dimension. Factor analysis also tends to treat responses x_{ij} as normally distributed which is frequently not the case; while this might be an acceptable approximation [Rhemtulla et al., 2012] in some scenarios, I worry that it is suboptimal for item banking. There is also a separate tradition of measurement invariance invoked in the factor analysis literature [Putnick and Bornstein, 2016]. IRT and factor analysis have relative strengths and weaknesses that perhaps suggest a sequencing of how they can be most effectively used. Exploratory work can be done using factor analysis to decompose complex attributes into distinct dimensions. Research can then build an appropriate probabilistic model that accounts for specifics of the response (e.g., these models are specific to the distribution of the response rather than being normal approximations). Such probabilistic models, being tailored to the response's distribution, allow for stringent empirical testing and more coherent item banking.

4.5 A quick summary

Let's rapidly review how an item bank built on probabilistic item response models can help resolve common measurement problems. The probabilistic model itself is falsifiable and allows for robust interrogation of whether it is an appropriate description of data. One particularly critical interrogation in this mode pertains to invariant functioning across groups; there are a range

[7] We can also consider differences between predictions of multidimensional and unidimensional models; they may be less pronounced than anticipated [Domingue et al., 2024].

of approaches that can be used to examine whether measurement invariance holds. Once affirmative empirical support exists for these propositions, the item bank can then be used to generate respondent estimates that are comparable across settings. Items drawn from the bank can be given in fixed form or, when respondent time is constrained, adaptively. Resulting ability estimates are always comparable as they are on the scale defined by the item parameters. When building item banks, probabilistic models are valuable tools for coherent understanding (via equating techniques) of responses generated via the use of different forms. These techniques offer efficient solutions to many of the most common measurement challenges faced in real-world settings and can be readily deployed to improve the practice of psychological measurement in a range of scenarios.

5 Examples

I will now discuss three measurement projects of relevance for developmental psychology with the goal of showcasing how they are using item-level understanding to address practical measurement problems. As appropriate, I will identify key challenges for such measurement systems with the aim of emphasizing the difficulties associated with developing such systems. Each of these systems is effectively developing an item bank which can be used to generate information on age-related change in the measured attributes. They all offer excellent opportunities for reuse (in the spirit of Elson et al. [2023]) given the extensive work that went into initially developing and calibrating these instruments.

5.1 NIH Baby Toolbox

The NIH Infant and Toddler Toolbox [Gershon et al., 2024] aims to measure neurodevelopment in young children (1–3.5y). It is similar to the widely-used NIH Toolbox [Gershon et al., 2013] which applies to older children. There are over 30 measures in the Baby Toolbox covering function in the domains of cognition, motor control, and social-emotional. It is focused on US children and is available in English and Spanish.

5.1.1 Norms based on a representative sample

Collecting data from a representative sample is costly in terms of time and money. Under assumptions of parameter invariance, it will not necessarily improve parameter estimates (although it will increase confidence that estimates are not varying across population subgroups) as compared to those derived from convenience samples. Given these facts, most psychological measures are built using data from convenience samples. The NIH Infant and Toddler Toolbox is a notable exception as it was calibrated based on a representative

sample of roughly 2500 children. While expensive, such samples are imperative if the goal is to understand the full distribution of developmental outcomes (i.e., norming).

Recruitment of a representative sample allows for calculation of age norms for the constructs. The aim is to offer context for the performance of a given child (e.g., given their age, this child's language skills are at the 70th percentile relative to some specific population). Given that a variety of measures in use in the Toolbox may be relevant for a broader construct (e.g., expressive and receptive vocabulary may both be relevant for understanding the development of language abilities), they combine information across measures to construct composite scores representative of a broader domain [Han et al., 2025b]. Age-related growth in these composites can be observed (see Figure 4 in Han et al. [2025a]); relatively little growth is observed for negative affect and self regulation as compared to the other measures. This approach wherein data from several measures is combined so that age-related changes can be studied will again appear when we discuss D-scores.

5.1.2 Considerations of statistical power

A second noteworthy feature of the Toolbox's calibration sample is their focus on a well-powered sample. Power considerations are a critical element in intervention studies; in measurement-related projects, they tend to play a less formal role. Given their interest in identifying developmental delays, the investigators considered a power analysis based on assumptions about a hypothetical measurement instrument (e.g., 75 items) being used to differentiate respondents in the left tail of the distribution. A calibration sample of 1000 respondents would have been sufficient for obtaining relatively accurate estimates around the average but would have produced relatively noisy estimates three SDs from the average; such considerations led to a desired sample size of around 2500 respondents. This focus on a well-powered representative sample is noteworthy; while the representative sample is probably outside what most research teams can achieve, a consideration of the power necessary for high-quality measurement is a practice that could be in more widespread use and is consistent with a broader interest in using power calculations to benchmark accuracy in parameter estimates [Maxwell et al., 2008].

5.2 LEVANTE

LEVANTE [Frank et al., 2025, Kachergis et al., 2025] is a measurement system focused on older children (2–12y). It is a suite of measures meant to be adapted to diverse settings. Pilot data were collected in Colombia, Canada, and Germany with new data being collected in additional countries in an ongoing fashion. LEVANTE contains both child direct assessments and caregiver reports; I will discuss the former.

5.2.1 Integration with Open Science

LEVANTE is designed to be consistent with open-science practices. While much of these details are not directly related to measurement, I do they think are worth noting especially given my interest in warehousing such data [Domingue et al., 2025]. The LEVANTE systems is structured such that de-identified data are available via permissive licensing terms in a centralized repository. This encourages both replication of original findings and reuse of data for novel investigations.

5.2.2 Measurement invariance

A key goal of LEVANTE is understanding variability in learning across a range of settings. Such work must grapple with a fundamental challenge as it necessitates having measures that function equivalently across settings. Absent that, claims regarding relative variation between sites would be meaningless. We thus require strong evidence regarding measurement invariance. This fundamental tension—trying to conduct measurement in a way that yields comparable measures across sites when possible but also being aware of the associated challenges and responsive to the fact that not all tasks will allow for such invariance—has been at the heart of the LEVANTE project.

As an illustration of the variation in task invariance, let us consider the LEVANTE instantiations of a Sentence Understanding task—based on the Test for Reception of Grammar (TROG; [Bishop, 1983]) which involves a child selecting a picture that matches a sentence which has been read aloud—and Vocabulary task—where a respondent matches a picture with a spoken word [Yeatman et al., 2021]—in the Colombian ($N = 368$ for sentence understanding and $N = 181$ for Vocabulary) and German samples ($N = 259$ for sentence understanding and $N = 264$ for Vocabulary).[8] To make things straightforward, we consider the raw proportions correct (i.e., $\overline{x_{.j}}$) for items across both samples (these $\overline{x_{.j}}$ are closely related to item difficulty b_j in Eqn 1). In the aggregate, the sentence understanding items that are difficult in one sample are expected to be difficult in the other (as measurement invariance would suggest) with a correlation being average proportions of 0.84. However, for Vocabulary, we only observe a correlation of 0.41 suggesting much weaker consistency in task difficulty. As a specific example, 33% of Colombian respondents identify 'tapestry' versus 79% in Germany. As with our discussion of 'horse' in Section 4.1, this is likely due to differences in the role of different words across the languages. While this is a simple descriptive analysis, it is indicative of a potent difference of the invariance of these instrument across contexts (see also Table 6 of Kachergis et al. [2025]).

What are the operational implications? In my view, these results offer some hope that the Sentence Understanding item parameters may be treated

[8] You can experiment with these tasks here: `https://researcher.levante-network.org/measures`.

as invariant across these groups thus allowing for between-group comparisons to be made. That said, differences in item performance can be monitored and investigated as to whether further refinements (e.g., removing items that seem to exhibit variation in function) lead to improved performance. Vocabulary, on the other hand, may require group-specific item parameters which will limit subsequent attempts to draw between-group comparisons. That's not to say that there is nothing to learn about between-group comparison of vocabulary—see Chapter 5 of [Frank et al., 2021]—but rather that comparisons must be made carefully and depending on specific words may be highly misleading.

5.3 D-score

The D-score [Weber et al., 2019] is designed to depict development across a range of international settings on a unidimensional interval scale for young children (< 36 months). While development of LEVANTE and the NIH Toolbox involved novel primary data collection, the D-score approach focused on building an item bank that links measures from different studies. The emphasis on unidimensionality is also in contrast to both the NIH Toolbox and LEVANTE. There are reasons we may want to have both fine-grained (as in the Toolbox) and more coarse measures of development. The Toolbox can be used as a diagnostic to better understand the developmental profile of a single child; its granularity will be useful in helping to identify specific areas of concern that may require intervention. Note that this granularity also comes at the price of contextual specificity (e.g., the Toolbox is only available in two languages). In contrast, the D-score can be rapidly used in a number of settings as a surveillance tool for understanding, for example, patterns of development that may be affected by policy changes.

5.3.1 Equating

At its heart, this is a large equating study which uses data from 16 longitudinal studies to create the D-score scale. How did this work? After some filtering, the authors are focusing on 565 items from the 16 studies which are taken from 11 instruments. Qualitative evaluation of the items yielded 95 'equate groups' which are groups of items that are effectively identical across instruments (e.g., 'stacks 2 cubes' in one instrument being in the same equate group as 'builds 2 block tower' in another). Additional evaluation (e.g., fit with the relevant item response model) reduced this to 18 active equate groups. Items within active equate group are considered to be common (or anchor) items across the instruments/cohorts. Specifically this means that item parameters—and they use the Rasch model in this analysis—are treated as equivalent for items within an equate group. The existence of responses from across studies to the common items allows for the mapping of results from all studies onto a common scale. The quality of this mapping hinges on assuming that these common items are representative of the full item set and that the the items in the active equate

groups are invariant across the contexts covered by the various studies (as well as in future/unobserved contexts given that they advocate for usage of the D-score scale in work moving forward).

5.3.2 Interval scale

The D-score is designed to represent an interval scale. Let us first discuss the unit. Scale values of 20 and 40 are fixed for a child being able to lift head to 45 degrees from the prone position and sitting in a stable position without support respectively. As a consequence, D-scores are near zero at one month and expected to increase by one unit each week (thus being around 44 when a child is 1y). Development slows in the second year; during this period, one unit increases are expected every month [Weber et al., 2019, p. 5-6]. Is the scale interval in the sense that, say, a D-score growth from 5 to 10 represent the same change in 'development' as growth from 65 to 70? It is difficult to be sure; the authors do examine consistency with the Rasch model but the Rasch ideal is not likely to be met and it is unclear what the implications of departures from the strong assumptions of the Rasch model are in terms of the resulting scale.

Where does this leave us? This is a theme that I will return to in the discussion but I make one observation here. Whatever the properties of the D-score scale, development also takes place across time. There is no ambiguity about time being intervally scaled. The authors make use of this to motivate the selection of unit—perhaps similar to the 'anchoring' approach used in economics [Cunha and Heckman, 2008]—but one could imagine extending this process to yield a 'developmental age' that instead places a child at some hypothesized age based on their observed capacities in a manner similar to the 'epigenetic age' [Duan et al., 2022] developed for aging research.

6 Discussion

Ideas of growth and change are at the heart of studies of human development and observations of such phenomena ultimately hinge on systems of measurement. While children develop physically, much of the key information for understanding subsequent well-being is psychological. Studies of child development thus necessitate psychological measures. I have tried to offer an overview of key ideas and principles relevant to such measures and have emphasized the utility of item banks as a shortcut to faster, more effective psychological measurement. The three empirical examples—the NIH Baby Toolbox, LEVANTE, and D-scores—are potentially useful exemplars of item banks that could be used to study a diverse array of attributes relevant across varying contextual ranges and ages. They are also designed for reuse. Given that I think developmental research may benefit from an increased emphasis on building item banks, I offered some specific suggestions for reporting on new

measurement approaches above. Below I emphasize several key points that might be relevant for researchers considering the design and usage of psychological measures before then turning to a discussion of some more challenging conceptual matters.

SUMMARY POINTS

1. Development of an item bank can help to resolve many measurement challenges. A functioning item bank will allow new respondents to be placed onto the same score scale as prior respondents which ensures comparability across time (under assumptions that item functioning is temporally invariant).

2. One key benefit of an item bank is that it can be utilized to facilitate adaptive testing. Adaptive testing can greatly reduce the time burden placed on respondents without sacrificing measurement precision.

3. Equating studies can be used to allow for comparability between scores produced via different forms. Use of 'common item' designs can be valuable in allowing for comparability between outcomes from two different measures that have some degree of overlap (i.e., the common items). Item banking also enables form construction using previously-established item parameters and thus eliminates the need for equating following data collection.

4. Measurement in heterogeneous samples is hard and concerns regarding measurement invariance should be at the forefront. These issues need to be managed proactively and, when serious, may necessitate separate scoring of respondents from different contexts or even reconceptualization of the measure. Item banks with information derived from one sample need to be scrutinized for invariance before they are used in diverse samples.

5. Measurement of physical attributes is a gold standard towards which measurement of psychological attributes can strive. We can rely on the features of physical measurement to help us identify desiderata of our systems for psychological measurement and we need to be sensitive about implications of failing to meet this standard.

I close with a final discussion on three important but contentious considerations starting with my claim that developmental psychology should attempt, when possible, to separately measure distinct unidimensional attributes, and then study their interplay, rather than working directly with multidimensional

measures. I suggest that researchers first identify attributes that can be separately measured via unidimensional scales and to subsequently use those *scales* to understand interplay between attributes rather than directly through the *items*. What is my rationale? I anticipate that much multidimensionality is 'between-item' rather than 'within-item'. When this is the case, we can more effectively study the intricacies of development across attributes based on coherent unidimensional measures rather than a bunch of items that needlessly tap multiple attributes.[9] Situations where we have to manage within-item multidimensionality will exist and may raise intriguing questions about the nature of the construct but should not be the goal of measure design.

Supposing that we can work with unidimensional measures, should we consider them as intervally scaled? Interval scaling should be a goal rather than an assumption. As in other scientific disciplines, we should be striving to build intervally scaled measures. Consider temperature: interval measures of temperature are a byproduct of a long sequence of scientific discoveries [Sherry, 2011]. Interval scales do not come easily! Our starting point should be one of curiosity about the structure of the attribute and whether this structure allows for interval measures. We should be in the business of hypothesizing that interval measures are possible for some attribute and then working to stringently test such a hypothesis; such testing can take the form of assessing the appropriateness of a specific item response model (especially relevant when working within the constraints of the Rasch models). Attributes that do not (yet?) allow for interval measurement will, of course, be identified and analyses that assess subsequent sensitivity of results would be required [Ballou, 2009]. Development of interval measures may be challenging but, as in other scientific fields, limitations of our measurement systems should be considered as fundamental limitations on our ability to understand the attributes of interest.

Let us end with validity. Debates about validity are long-standing and I think that the particular debate about the role of measurement 'uses' and the degree to which assessments of validity should hinge on such uses is intractable (and, frankly, perhaps telling of fundamental difference between lower-stakes psychological measures and higher-stakes educational measures). I shall leave be the Gordian knot at the heart of this debate and instead encourage a focus on *psychological measurement as measurement*. A heightened scrutiny on the predictive accuracy of the relevant measurement models and whether we can exert precise control over that accuracy by changing features of items,

[9]We can (and do) separately measure height and weight as kids develop and use those measures to understand physical growth. Imagine that, instead, we used longitudinal information from a bunch of different indicators that pooled information about overall body size—Does the child fit in 3 month clothes? Do you have a hard time fitting their head through neck of shirts? What is their shoe size? What size car seat do they require?—and tried to base our understanding of their length and mass on that. The shortcomings of such an approach are apparent yet this is what we are often doing with psychological measurement.

for example, will push psychological measurement in interesting directions. More generally, focused scrutiny on the relevant developmental changes that can be assessed by a measure, the degree to which performance of a measure is invariant across context, establishment of item banks that allow for more rapid data collection for vetted measures: these are problems that can be addressed using some of the tools discussed here and from which even partial improvements will benefit developmental psychology.

DISCLOSURE STATEMENT

The author is affiliated with the team developing LEVANTE.

ACKNOWLEDGMENTS

The author would like to thank Emma Armstrong-Carter, Paige Harden, Klint Kanopka, Mateus Mazzaferro, Sanford Student, and Lijin Zhang for helpful feedback on earlier drafts. Drew Bailey also read it.

References

David Andrich. A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling thurstone and likert methodologies. *British Journal of Mathematical and Statistical Psychology*, 49(2):347–365, 1996.

Frank B Baker and Seock-Ho Kim. *Item response theory: Parameter estimation techniques*. CRC press, 2004.

Dale Ballou. Test scaling and value-added measurement. *Education finance and Policy*, 4(4):351–383, 2009.

Daniel J Bauer and Patrick J Curran. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological methods*, 8(3):338, 2003.

William Belzak and Daniel J Bauer. Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological methods*, 25(6):673, 2020.

DV Bishop. Test for the reception of grammar, age and cognitive performance; research centre. *University of Manchester: Manchester, UK*, 1983.

R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4): 443–459, 1981.

R Darrell Bock and Michele F Zimowski. Multiple group irt. In *Handbook of modern item response theory*, pages 433–448. Springer, 1997.

R Darrell Bock, Eiji Murakl, and Will Pfeiffenberger. Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25(4):275–285, 1988.

Kenneth A Bollen. Latent variables in psychology and the social sciences. *Annual review of psychology*, 53(1):605–634, 2002.

Wes Bonifay and Li Cai. On the complexity of item response theory models. *Multivariate behavioral research*, 52(4):465–484, 2017.

Elaine Borghi and Harshpal Singh Sachdev. Should a single growth standard be used to judge the nutritional status of children under age 5 y globally? debate consensus. *The American Journal of Clinical Nutrition*, 120(4):769–772, 2024.

Denny Borsboom, Gideon J Mellenbergh, and Jaap Van Heerden. The concept of validity. *Psychological review*, 111(4):1061, 2004.

Derek C Briggs. Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2):204–226, 2013.

Derek C Briggs. *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge, 2021.

David Budescu. Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement*, 22(1):13–20, 1985.

Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540, 1992.

Stephanie Buono and Eunice Eunhee Jang. The effect of linguistic factors on assessment of english language learners' mathematical ability: A differential item functioning analysis. *Educational Assessment*, 26(2):125–144, 2021.

Gregory Camilli. Test fairness. *Educational measurement*, 4:221–256, 2006.

R Philip Chalmers. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71:1–38, 2016.

Hua-Hua Chang. Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1):1–20, 2015.

Gregory L Cizek and Robert J Sternberg. *Setting performance standards*. Lawrence Erlbaum Associates Mahwah, NJ, 2001.

Flavio Cunha and James J Heckman. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of human resources*, 43(4):738–782, 2008.

Clintin P Davis-Stober. Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, 53(1):1–13, 2009.

Adele Diamond. Executive functions. *Annual review of psychology*, 64(1): 135–168, 2013.

Ben Domingue. Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, 79(1):1–19, 2014.

Benjamin W Domingue, Ryan J McCammon, Brady T West, Kenneth M Langa, David R Weir, and Jessica Faul. The mode effect of web-based surveying on the 2018 us health and retirement study measure of cognitive functioning. *The Journals of Gerontology: Series B*, 78(9):1466–1473, 2023.

Benjamin W Domingue, Klint Kanopka, Radhika Kapoor, Steffi Pohl, R Philip Chalmers, Charles Rahal, and Mijke Rhemtulla. The intermodel vigorish as a lens for understanding (and quantifying) the value of item response models for dichotomously coded items. *psychometrika*, 89(3):1034–1054, 2024.

Benjamin W Domingue, Mika Braginsky, Lucy Caffrey-Maffei, Joshua B Gilbert, Klint Kanopka, Radhika Kapoor, Hansol Lee, Yiqing Liu, Savira Nadela, Guanzhong Pan, et al. An introduction to the item response warehouse (irw): A resource for enhancing data usage in psychometrics. *Behavior Research Methods*, 57(10):1–11, 2025.

Ran Duan, Qiaoyu Fu, Yu Sun, and Qingfeng Li. Epigenetic clock: A promising biomarker and practical tool in aging. *Ageing research reviews*, 81: 101743, 2022.

Michael C Edwards, Carrie R Houts, and Li Cai. A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological methods*, 23(1):138, 2018.

Malte Elson, Ian Hussey, Taym Alsalti, and Ruben C Arslan. Psychological measures aren't toothbrushes. *Communications Psychology*, 1(1):25, 2023.

Susan E Embretson. The new rules of measurement. *Psychological assessment*, 8(4):341, 1996.

A. Ferguson, C. S. Myers, R. J. Bartlett, H. Banister, F. C. Bartlett, W. Brown, and W. S. Tucker. Quantitative estimates of sensory events, final report, 1940.

Gerhard H Fischer and Ivo W Molenaar. *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, 1995. ISBN 978-1-4612-4230-7.

Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694, 2017.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. *Variability and consistency in early language learning: The Wordbank project*. MIT Press, 2021.

Michael C Frank, Heidi A Baumgartner, Mika Braginsky, George Kachergis, Amy A Lightbody, Robert Z Sparks, Rebecca Zhu, Stephanie M Carlson, Sandra Graham, Sebastián J Lipina, et al. Learning variability network exchange (levante): A global framework for measuring children's learning variability through collaborative data sharing. *Child Development*, 2025.

Richard Gershon, Miriam A Novack, and Aaron J Kaat. The nih infant and toddler toolbox: A new standardized tool for assessing neurodevelopment in children ages 1–42 months. *Child development*, 95(6):2252–2254, 2024.

Richard C Gershon, Molly V Wagster, Hugh C Hendrie, Nathan A Fox, Karon F Cook, and Cindy J Nowinski. Nih toolbox for assessment of neurological and behavioral function. *Neurology*, 80(11_supplement_3):S2–S6, 2013.

Sasan Ghorbani, Seyed Hadi Hoseinie, Ebrahim Ghasemi, Taghi Sherizadeh, and Christina Wanhainen. A new rock hardness classification system based on portable dynamic testing. *Bulletin of Engineering Geology and the Environment*, 81(5):179, 2022.

Joshua B Gilbert, Benjamin W Domingue, and James S Kim. Estimating causal effects on psychological networks using item response theory. *Psychological Methods*, 2025.

Peter F Halpin. Differential item functioning via robust scaling. *psychometrika*, 89(3):796–821, 2024.

Y Catherine Han, Elizabeth M Dworak, Maxwell Mansolf, Hubert Adam, Lihua Yao, Miriam A Novack, Sarah Pila, Rachel M Flynn, Amanda M Flagg, Vitali Ustsinovich, et al. Nih baby toolbox® methodology and norms development. *Infant Behavior and Development*, 80:102117, 2025a.

Y Catherine Han, Elizabeth M Dworak, Maxwell Mansolf, Richard C Gershon, and Aaron J Kaat. Composite scores for the nih baby toolbox®. *Infant Behavior and Development*, 80:102122, 2025b.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Joint Committee on the Standards for Educational and Psychological Testing. *Standards for educational and psychological testing*. American Educational Research Association, Washington DC, 2014.

George Kachergis, Fionnuala O'Reilly, Mika Braginsky, Xingyao Xiao, Amy Lightbody, KA Shannon, Zachary Watson, Lijin Zhang, Rebecca Zhu, AB Abutto, et al. Creation and validation of the levante core tasks: Internationalized measures of learning and development for children ages 5-12 years. 2025.

George Karabatsos. The rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of applied measurement*, 2(4):389–423, 2001.

George Karabatsos. On bayesian testing of additive conjoint measurement axioms using synthetic likelihood. *psychometrika*, 83(2):321–332, 2018.

Michael J Kolen and Robert L Brennan. *Test equating: Methods and practices*. Springer Science & Business Media, 2013.

David H Krantz, Patrick Suppes, and R Duncan Luce. *Additive and polynomial representations*, volume 1. Courier Corporation, 2006.

Samuel A Livingston. Equating test scores (without irt). *Educational testing service*, 2014.

Frederic M Lord. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum, Hillsdale, NJ, 1980.

Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. Addison-Wesley, 1968.

R Duncan Luce and John W Tukey. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, 1(1):1–27, 1964.

Wanjing Anya Ma, Yiqing Liu, Wenchao Ma Klint Kanopka, and Benjamin W Domingue. A comparison of the predictive performance of continuous and class-based latent trait models. 2025a.

Wanjing Anya Ma, Adam Richie-Halford, Amy K Burkhardt, Klint Kanopka, Clementine Chou, Benjamin W Domingue, and Jason D Yeatman. Roar-cat: Rapid online assessment of reading ability with computerized adaptive testing. *Behavior Research Methods*, 57(1):56, 2025b.

Esther Maassen, E Damiano D'Urso, Marcel ALM Van Assen, Michèle B Nuijten, Kim De Roover, and Jelte M Wicherts. The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*, 2023.

David Magis and Gilles Raîche. Random generation of response patterns under computerized adaptive testing with the r package catr. *Journal of Statistical Software*, 48:1–31, 2012.

David Magis, Sebastien Beland, Francis Tuerlinckx, and Paul De Boeck. A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior research methods*, 42(3):847–862, 2010.

Luca Mari. A quest for the definition of measurement. *Measurement*, 46(8): 2889–2895, 2013.

Scott E Maxwell, Ken Kelley, and Joseph R Rausch. Sample size planning for statistical power and accuracy in parameter estimation. *Annu. Rev. Psychol.*, 59(1):537–563, 2008.

Joel Michell. Quantitative science and the definition of measurement in psychology. *British journal of Psychology*, 88(3):355–383, 1997.

Joel Michell. *Measurement in psychology: A critical history of a methodological concept*, volume 53. Cambridge University Press, 1999.

Joel Michell. Measurement: overview. *Wiley StatsRef: Statistics Reference Online*, 2014.

Roger E Millsap. *Statistical approaches to measurement invariance*. Routledge, 2012.

Victoria J Molfese, Peter J Molfese, Dennis L Molfese, Kathleen Moritz Rudasill, Natalie Armstrong, and Gillian Starkey. Executive function skills of 6–8 year olds: Brain and behavioral evidence and implications for school achievement. *Contemporary educational psychology*, 35(2):116–125, 2010.

Michael Nering and Remo Ostini. *Handbook of polytomous item response theory models*. Taylor & Francis, 2011.

Richard Perline, Benjamin D Wright, and Howard Wainer. The rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3 (2):237–255, 1979.

Diane L Putnick and Marc H Bornstein. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, 41:71–90, 2016.

Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Mark D Reckase. 18 multidimensional item response theory. *Handbook of statistics*, 26:607–642, 2006.

Hao Ren, Wim J van der Linden, and Qi Diao. Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika*, 82(2): 498–522, 2017.

Mijke Rhemtulla, Patricia É Brosseau-Liard, and Victoria Savalei. When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological methods*, 17(3):354, 2012.

Joseph P Robinson. Evaluating criteria for english learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 33(3):267–292, 2011.

David Sherry. Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, 42(4):509–524, 2011.

Hyejin Shim, Wes Bonifay, and Wolfgang Wiedermann. Parsimonious item response theory modeling with the cauchit link: Revisiting the rationale of the four-parameter logistic model. *Behavior Research Methods*, 57(6):176, 2025.

Klaas Sijtsma, Jules L Ellis, and Denny Borsboom. Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika*, 89(1): 84–117, 2024.

Guillermo Solano-Flores, Carne Barnett-Clarke, and Rachel R Kachchaf. Semiotic structure and meaning making: The performance of english language learners on mathematics tests. *Educational assessment*, 18(3):147–161, 2013.

Benjamin A Stenhaug and Benjamin W Domingue. Predictive fit metrics for item response models. *Applied Psychological Measurement*, 46(2):136–155, 2022.

A Jackson Stenner, Hal Burdick, Eleanor E Sanford, and Donald S Burdick. How accurate are lexile text measures? *Journal of Applied Measurement*, 7 (3):307, 2006.

Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103 (2684):677–680, 1946.

Martha L Stocking and Frederic M Lord. Developing a common metric in item response theory. *Applied psychological measurement*, 7(2):201–210, 1983.

William Stout. A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4):589–617, 1987.

Yoshio Takane and Jan De Leeuw. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3): 393–408, 1987.

David Torres Irribarra. A pragmatic perspective of measurement. In *A Pragmatic Perspective of Measurement*, pages 43–62. Springer, 2021.

Gerhard Tutz. A taxonomy of polytomous item response models. *arXiv preprint arXiv:2010.01382*, 2020.

C David Vale. Computerized item banking. *Handbook of test development*, pages 261–285, 2006.

Wim J Van der Linden and Seung W Choi. Improving item-exposure control in adaptive testing. *Journal of educational measurement*, 57(3):405–422, 2020.

Wijbrandt H Van Schuur. Mokken scale analysis: Between the guttman scale and parametric item response theory. *Political Analysis*, 11(2):139–163, 2003.

Matthias von Davier and Young-Sun Lee. Handbook of diagnostic classification models. *Cham: Springer International Publishing*, 2019.

Howard Wainer, Eric T Bradlow, and Xiaohui Wang. *Testlet response theory and its applications*. Cambridge University Press, 2007.

Ann M Weber, Marta Rubio-Codina, Susan P Walker, Stef Van Buuren, Iris Eekhout, Sally M Grantham-McGregor, Maria Caridad Araujo, Susan M Chang, Lia CH Fernald, Jena Derakhshani Hamadani, et al. The d-score: a metric for interpreting the early development of infants and toddlers across global settings. *BMJ global health*, 4(6):e001724, 2019.

Benjamin D Wright. A history of social science measurement. *Educational measurement: Issues and practice*, 16(4):33–45, 1997.

Margaret Wu and Richard J Adams. Properties of rasch residual fit statistics. *Journal of Applied Measurement*, 14(4):339–355, 2013.

Jason D Yeatman, Kenny An Tang, Patrick M Donnelly, Maya Yablonski, Mahalakshmi Ramamurthy, Iliana I Karipidis, Sendy Caffarra, Megumi E Takada, Klint Kanopka, Michal Ben-Shachar, et al. Rapid online assessment of reading ability. *Scientific reports*, 11(1):6396, 2021.