Measuring Unobserved Variables in Educational Inequality Research: Mathematics and Language 'Abilities' in Early Childhood

Alejandra Rodriguez S.

July 18, 2019

Abstract

Evidence of social inequalities in cognitive abilities in early childhood has been documented in many societies; however, three characteristics of the data used to measure cognitive constructs make it difficult to quantify inequalities across groups. First, a causal understanding of validity is not compatible with the standard validation framework, which forces researchers to think critically what it means to measure unobserved constructs. Second, test scores only provide ordinal information about individuals, they are not interval scales and require the use of suitable corresponding methods for their study. Third, measurement invariance, which causes measurement error, may make comparison of test scores across groups invalid. The paper explores these three data problems applied to standardized tests—one mathematics and two language assessments—taken by a cohort of German children. The paper proposes a comparative validation framework for researchers based on nonparametric psychometric models and the representational theory of measurement. This framework can help researchers to determine if data fit the assumptions of a measurement model, to check for various forms of measurement error, and to overcome potential issues. A comparison of competing statistical modeling alternatives reveals substantial differences: By conceptualizing ability as ordinal instead of interval and excluding items that do not fit the assumptions of measurement models, I find a reduction in effect sizes for typical covariates studied in social stratification research.

A number of studies—from Noble et al. (2015), Kalil, Ryan, and Corey (2012), Duncan, Ziol-Guest, and Kalil (2010), Halle et al. (2009) among others—have shown that social inequalities in cognitive abilities appear early on in life. Blossfeld, Kulic, and Triventi (2017, 89–105), Weinert et al. (2016) and Solga and Dombrowski (2009) have shown similar inequalities among German children. Take, for instance, the probability of scoring below the 25th percentile of the distribution of mathematics ability in a standardized mathematics test (Petersen and Gerken (2018), Blossfeld, Roßbach, and Maurice (2011)). For preterm children, the relative risk of being among this lowest scoring group is 1.724 times that of fullterm babies (C.I: [1.323, 2.208]); for girls, the risk is 0.979 times that of boys (C.I: [0.855, 1.121]); for children with migration background it is 1.736 times that of nonmigrant background children (C.I: [1.45, 2.063]); and for children of parents in the least well off socioeconomic status (SES) 3.162 times the risk than children of the most wealthy, highly educated and better employed parents (C.I: [2.569, 3.902]); own calculations). Such inequalities have been shown to exist in developed and developing countries alike (the US: Breda, Jouini, and Napp (2018); the UK: Schoon (2010) and Ermisch (2008); France: Jednoróg et al. (2012) or Goussé and Le Donné (2014); Brazil: Tella et al. (2018) and Wehby and Trujillo (2017) ; Colombia: Gamboa and Londoño (2015); etc.).

The main reason why researchers are concerned about early life educational inequalities is that they may persist over time through still not fully understood mechanisms such as: cumulative disadvantages (Lee and Jackson (2017)); dynamic complementarities (Cunha, Heckman, and Schennach (2010)); Mathew effects (Protopapas, Parrila, and Simos (2016)); social status dominance (Van Laar and Sidanius (2001)); and institutional discrimination through teacher's biased assessments (Batruch et al. (2019), Croizet and Dutrévis (2004) and Millet and Croizet (2016)). Recent research has focused on overcoming these inequalities through pre-K or early childhood interventions, given that effects of school-age alleviation programs have been shown to fade over time (Becker (2011), Heckman, Humphries, and Kautz (2014, 341–430)). Policies targeting infants and toddlers cognitive capacities are justified based on the estimated higher returns to investments in early childhood. The first years of life are fundamental to children's wellbeing as adults, as summarized by the human capital slogan: "skills beget skills" (Heckman (2006)); moreover, neuroscientific evidence shows differences by SES are present in the developmental structures and functions of children's brains (Farah (2017), S. B. Johnson, Riis, and Noble (2016)), further substantiating the need for early childhood interventions.

However, the characteristics of the data used to measure cognitive constructs troubles the quantification of social inequalities in cognitive abilities as measured by test scores in standardized tests, as well as the study of their underlying causal and explanatory mechanisms, and also the evaluation of interventions to overcome these inequalities.

Background

First, causal understandings of validity are not compatible with the traditional validation framework. "A test is valid for measuring an attribute if variation in the attribute causes variation in the test scores" (Borsboom, Mellenbergh, and Heerden (2004, 1067)). Therefore, changes in brain structure and function at a neuronal level that correspond to higher cognitive ability should cause nonnegative changes in test scores if researchers are to conclude that a standardized test is valid, i.e., that it measures something happening to the brain (e.g. more neural connections leading to higher mathematics ability, similar to a blood count test where the presence of disease causes changes in the cell count). In the psychometric framework of item response theory (IRT), cognitive abilities—for example, mathematics ability—together with the structures and functions of the brain indicative of a more mathematically able child are presumed to lie on a $[0, max]$ range; where 0 corresponds to no ability and max corresponds to maximum mathematics ability. Test scores, which have their own arbitrary scales, should map onto the hypothesized range of mathematics ability in the brain in a nondecreasing functional relation (Vautier et al. (2012)). Such evidence has not been provided by test developers. Aside from studies that find that young infants possess fundamental cognitive capacities to differentiate auditory and visual stimuli (G. Dehaene-Lambertz and Spelke (2015)), standardized tests aimed at "measuring" cognitive abilities use the verb measure in a metaphorical sense (Briggs (2013)).

Even though no unique concept of validity exists, as evidenced in discussions around "The great validity debate" (S. B. Johnson, Riis, and Noble (2016)), validity is still downplayed in empirical social research, where test scores are used routinely to show gaps and explain gaps. Present state-of-the-art methods follow the *Standards for Educational and Psychological Testing* (A. E. R. Association, Association, and Measurement in Education (2014)) to assess the internal consistency of latent variables. This is estimated through manifest responses to Likert or polytomous type items and the sum of correct answers, which is sometimes scaled by an IRT model. The validation strategy for scales of psychological constructs follows the traditional validation framework as surveyed in Shear and Zumbo (2014, 91–111) where a combination of exploratory or confirmatory factor analysis (EFA or CFA) procedures, internal reliability estimates such as Chronbach's α, correlation coefficients among test scores and other relevant outcomes are used. This is how validity claims are usually grounded.

Nevertheless, this understanding of validity might be entirely misleading. As shown by Maul (2017) and earlier by Wood (1978) among others, CFA or IRT do not provide researchers with the means of detecting a true underlying structure. These procedures cannot falsify the hypothesis that an underlying structure is driving the correlations, because almost always the methods will find some "structure" when applied to data; they are, after all, dimensionality-reduction techniques based on correlation, not causation. Moreover, Rasch models and its developments (2PL, 3PL, graded, etc.) have been shown to provide appropriate fit even when their assumptions in simulation studies are violated (see Karabatsos (2001, 394–95) for these results and Wood (1978) for estimating “coin flipping” ability). The construct of gibberish can be measured with high reliability mainly because these numerical procedures are not connected to the most intuitive definition of validity, namely that a standardized test measures what is supposed to measure and nothing else (Borsboom, Mellenbergh, and Heerden (2004)).

Second, the level at which researchers can measure such cognitive constructs is still a matter of dispute. The debate around the type of scale that a test scores is, whether ordinal or metric, remains relevant, regardless of how strongly scales of one kind are deemed to correlate with scales of the other. O’Brien (1985) provided an overview of the problems associated with transformation, categorization and grouping errors that result from using ordinal variables as interval ones. Treating an ordinal dependent variable as an interval scale may lead to several misleading results in the estimation of effects, such as arbitrariness of the scores or lack of cardinality, prediction values below or above the scale’s range, a lack of variability in response options that is subsequently ignored, ceiling and floor effects, among others (Agresti (2012, 5–7)). Additionally, Tymothy N. Bond and Lang (2013) discussed the importance of the assumptions underlying the scoring process of achievement tests. If test scores only have ordinal properties, then different assumptions about the unknown distribution of the unobserved latent construct (i.e. ability as not normally distributed), all of which may be valid, can lead to different results: a narrowing, a growing or a constant achievement gap in mathematics between Black and White students in the US. By assuming other distributions for the hypothesized mathematics ability, which can be done by applying nondecreasing monotonic transformations of the scores, dramatic changes arise in the direction of effects. In the value-added assessment literature, the same problem of treating an ordinal variable as an interval one has been noted (Ballou (2009)). Recently, Liddell and Kruschke (2018) enumerated several systematic errors that occur when doing so: false alarms, failure to detect effects and inversion of effects. Some of these difficulties have been discussed in the sociology of education literature (e.g. Protopapas, Parrila, and Simos (2016) and Nash

(2001)), yet applied researchers and policymakers continue to draw conclusions from methods that assume equal distances between units on an ordinal scale. The association between an ordinal dependent variable and another variable in general cannot be discerned, when the former is treated as an interval, i.e. when it is treated as if if had a quantitative structure. Developmental psychologists have yet to provide conclusive evidence that constructs assessed through standardized tests, such as the Peabody Picture Vocabulary Test Fourth Edition (PPVT-4) or other general or specific cognitive or noncognitive tests, are quantitative and that these constructs have been measured on an interval scale.

Third, serious epistemological criticisms of the use of psychometric scaling models to measure cognitive constructs abound. For an overview of literature discussing psychometrics as an entirely or partly pathological science, see H. Johnson (1936), Michell (1997), Michell (2008), Trendler (2013), Trendler (2009), Humphry (2013), Mari et al. (2017), Maul, Irribarra, and Wilson (2016), Briggs (2013), Vautier et al. (2012) and E. Lacot, Afzali, and Vautier (2016), among others. In a nutshell, these criticisms relate to two characteristics of standardized tests. First, they note that questionnaires or tests make explicit use of the human mind for measurement, i.e. a child answers questions on a test; and second, they point out that the human mind is not a reliable measuring device, i.e. the child might be distracted, unmotivated, bored or worried by other things, regardless of the control over the testing situation. In standardized educational tests, correct responses allow observers to infer that a child masters a particular skill or competence (excluding the possibility of guessing, in which case the inference is not warranted); however, the opposite inference cannot be drawn from an incorrect response, because a wrong answer might have resulted from language barriers; unfamiliarity with the testing situation; lack of concentration; lack of motivation; lack of confidence; stereotype threat; tiredness; stress, anxiety, or fear; an interaction of these factors; or from not knowing the correct answer (see Banerjee (2016) for a systematic review of factors affecting students' performance on standardized mathematics and science tests). Beyond traditional factors accounting for performance in educational research (e.g. opportunities to learn as determined by teacher, classroom or socioeconomic characteristics), extraneous factors that hinder students' performance are far from randomly distributed among a population; in fact, they are likely concentrated among disadvantaged groups. Moreover, they cannot be factored out from the measurement operation. It would be easy to derive hypotheses to explain differences or changes in scores as not unequivocally caused by corresponding differences or changes in the hypothetical attribute residing in the brain, which make tests invalid for establishing differences or inferring changes in psychological or educational constructs.

Measurement error in standardized tests complicates comparisons across social groups (see Boyd et al. (2013) for a general discussion of measurement error from an alternative perspective). Measurement invariance (MI), which is as a source of measurement error related to the above problems, means tests do not function equally for different groups in a population; some aspect of the test or the item affects the response behavior in a manner not relevant to the construct. When comparing groups, such as groups based on gender, race or ethnicity, and social class, tests must perform equally for all. Test developers have stated that pilot studies on similar samples allow the detection and exclusion of items displaying differential item functioning (DIF; Penfield and Camilli (2006)). Nevertheless, even within the traditional framework, validation is understood as an ongoing process (Chan (2014, 4)). A scale's demonstrably "good" psychometric properties in similar samples is not sufficient criteria to determine validity. A test is valid if, and only if, it has been validated for well-defined purposes in the contexts of its application. No DIF should be observed each time the test is used when comparing groups.

Psychologist's "atomic bomb" (to coin a term used by Borsboom and Wijsen (2017) referring to psychological tests) surreptitiously entered social science research practices without causing too much noise, but perhaps a lot of damage. Following the work of Baird et al. (2017), and recommendations made by Borsboom and Wijsen (2017), the paper proposes a comparative validation framework that tries to overcome the pitfalls of present validation practices in educational research by grounding the validation process in empirically supported claims. The proposal is based on methods within psychometrics (i.e. nonparametric item response theory (NIRT) and DIF) and the representational measurement theory (i.e. additive conjoint measurement). Such a validation framework should serve as a guarantee for the kinds of claims researchers, educators and policymakers are willing to make when interpreting test scores.

The paper presents a statistical and psychometric assessment of standardized cognitive ability tests (but applying to all inherent, unobservable, underlying, latent constructs—whether psychological, social and political—and whether measured in children, teenagers or adults) with two perspectives in mind. First, Tymothy N. Bond and Lang (2013) and Timothy N Bond and Lang (2014) criticisms have undermined empirical analysis of scales that lack cardinality, because there is no statistical or biological reason to assume that the distribution of an unobserved trait follows the "bell-curve". Before making comparisons, researchers need to establish whether their scales are interval or ordinal or if they at least allow for the classification of individuals into nominal groups (as originally intended by Alfred Binet, the founder of IQ testing; see Michell (2012), and see H. Johnson (1936) for pointing out the flaws of psychological measurement endeavors, still pertinent to this day). And second, following a pragmatist and social-construcivist sociology of education (Baird et al. (2017), Adams (2006)), concerned with "democratic equality" in educational opportunities (Elizabeth (2017)), it is crucial to solidly establish whether standardized tests display DIF among groups of interest or not.

The aim of the paper is to evaluate the psychometric properties of a mathematics test, and, to a lesser extent, those of two other widely used standardized psychological tests of language ability in children. The paper will also seek to establish if their use as an explanatory factor in social stratification research is justified. To do so, the paper proposes a framework for assessing the validity of standardized tests used to determine social inequalities. A standardized test and their resulting test scores are taken as prime examples of social constructions; not that the word latent *construct* would have made researchers think otherwise.

Methods

Assessing the development of children requires researchers to considerate multiple dimensions, such as language, socioemotional, behavioral, reasoning or cognitive, etc. The National Education Panel Study Starting Cohort No. 1 (NEPS SC1; Blossfeld, Roßbach, and Maurice (2011)) contains a diverse set of widely used standardized tests to measured such dimensions. These are a mixture of self-report items (by mostly mothers) and observational variables (collected by trained staff during field work). This section describes the main characteristics of three such standardized tests. The section presents the data analysis in a stepwise manner and describes the characteristics of the sample used for the analysis.

Participants

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Newborns, doi:10.5157/NEPS:SC1:6.0.0 . From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. The SC1 NEPS sample consists of a cohort of newborns in Germany that has been followed for over seven consecutive years. The sample was generated using a complex random study design and consists of officially registered newborns collected between February and July of 2012 (see Blossfeld, Roßbach, and Maurice (2011), Bauer (2014) for descriptions of the sampling mechanism). In each follow-up survey, the children's characteristics and performance in standardized tests were recorded along with characteristics of their parents, their households and their care facilities. The fifth Wave contains information for 2381 (49% girls) with median age of 51 months and range 47-56 months. This sample may still be considered as random, since attrition at the fifth Wave did not show considerable selectivity patterns (see Zinn et al. (2018) for a report of these studies).

Measurement Instruments

The paper's focus is on the following three standardized tests. A standardized **Mathematics test** in Wave 5 children was carried (Petersen and Gerken (2018)). This test consisted of 20 items on five different categories of mathematics competency: a) sets, numbers, and operations; b) units and measuring; c) space and shape; d) change and relationships; and e) data and chance. The **PPVT-4 Verbal ability** is the standard PPVT-4 (German adaptation) and was applied at Wave 4. The PPVT-4 assesses receptive vocabulary or verbal skills in children and adolescents (from 3 to 16 years of age in the German norm-referenced sample, see Roßbach, Tietze, and Weinert (2005)). The test contains 228 items divided into 19 sets, each of 12 items. Items in the PPVT-4 are of varied difficulty; however the total number of correct responses is computed to establish progression of children's language development. The PPVT-4 is also used as an indicator of cognitive ability, given that comprehension of language and concepts plays a major role in it. Sufficient reliability estimates are provided by the test developers. Finally, the **ELFRA-2 productive subscale (ELFRA-2P)** is The Parent's Questionnaire for the Early Diagnostic of Children at Risk 2 (Elternfragebogen fÃ¼r die FrÃ¼herkennung von Risikokindern 2; ELFRA-2 for name in German) and was assessed at Wave 3. The ELFRA-2 is a questionnaire filled in by parents of the child and consists of three scales that assess productive vocabulary (260 items) in the German language. The syntactic (64 items) and morphological (11 items) subscales were not considered here. The ELFRA-2 is used as a screening-test for diagnosing delays in language development in German children aged 24 months (Grimm and Doil (2006)). The scale has been shown to have appropriate reliability values (see Sachse and Von Suchodoletz (2007), Sachse and Suchodoletz (2007)), but is sensitive to the language spoken at home.

Data analysis

I have organized the data analysis in a step-wise manner in the hope that other researchers will put their data on standardized tests to the test. First, the paper presents traditional validation procedures and applies them to the mathematics test data only. The paper proposes the mini-test challenge as a measure of internal validity in the presence of potential sources of measurement invariance. It then applies additive conjoint measurement checks to assess the hypothesis of quantifiability of the attributes mathematics, language production and verbal ability. An assessment of ordinal scale properties through Mokken scale analysis follows; this again focuses on the mathematics test only for illustration purposes. From the results of those checks, the paper presents a comparison of different analysis strategies. Finally, measurement invariance in the three standardized tests is assessed.

Researchers typically design standardized tests with items of varying difficulty to assess a particular underlying construct. When the difficulty of items differs, raw or standardized sum scores do not yield an interval scale (Wright (1992) and Ballou (2009) discuss this uncontroversial point); sums provide only ordinal information that allows researchers to rank children. For this reason, researchers take care when scaling the data with IRT models; these models are supposed to guarantee that estimates of ability are on an interval scale. The literature on psychometrics as a pathological science argues that such assumptions are problematic, and moreover that, no scale has been found to be interval (for more criticisms from the perspective of mathematical psychology see Kyngdon (2010) and Domingue (2014); and for recent arguments that psychological attributes cannot be measured at all, see Trendler (2018)). Cumulative link models (CLM) are better suited to model ordinal cognitive constructs. This statistical model may be motivated from the same idea of latent construct. Winkelmann and Boes (2006, 175) have shown that the relationship between the scales $y_{[i]}$ (score on a test for child $i = 1, ..., N$) and the unobserved latent variable $\theta_{[i]}$ (underlying "ability" on $[0, max]$ range for child i) may be modeled using a threshold mechanism: $y_{[i]} = j$ if and only if

$\kappa_{j-1} < \theta_{[i]} \leq \kappa_j$ for $j = 1, ..., J$ correct responses, or total score with $0 < 1 < 2 < ... < J$ raw scores, consisting then of $J + 1$ unknown threshold parameters $\kappa_0 < ... < \kappa_J$ that partition the distribution of the unobserved latent trait. Therefore, assuming that a test score is only providing ordinal information still allows for empirical studies of social inequalities without adding unrealistic assumptions.

The present analytic strategy proposes three steps to overcome this disagreement. This involves, first, testing if standardized tests fit the assumptions of a quantitative structure, which should in principle be scalable by the IRT Rasch model. When these assumptions are not met, other assumptions of psychometric models might hold. Second, these assumptions are in turn tested by means of Mokken scale analysis (MKS; Sijtsma and Meijer (2006)). In this study, I present the checks for the properties of Mokken scales for the mathematics test, but in principle these can be checked for other standardized tests with a large number of items. Third, a comparison between modeling alternatives is presented: linear regression estimates using IRT ability estimates are compared with sums of scores or with a treatment of these estimates or sums as ordinal variables using cumulative link models.

The analysis starts by using the methods within the traditional validation framework:

Step 1: Explore items by looking at percentage of missing observations per item. If an item is deemed to have too large a percentage of missing responses with a cutoff of 50%; or too little variability, when more than 95% of responses are in a single category or were correctly or incorrectly answered providing little information; the item should be dropped from the analysis, entirely or at least at an initial stage.

Step 2: Compute reliability estimates, beginning by estimating association among items. For this purpose, the polychoric correlation is an appropriate statistic to assess associations among ordinal and binary variables. For reliability, coefficients α, Guttman's λ_2, and hierarchical ω_h are computed (for an overview of the salience of these other metrics versus the more widely used Chronbach's α see Revelle and Zinbarg (2009)). In addition, inter-item correlations as well as drop-item reliabilities are computed to examine internal reliability.

Step 3: Explore the underlying structure by computing EFA, CFA and structural equation models (SEM). For the EFA, I chose a larger number of hypothesized factors present in the mathematics test in order to compare the fit of different solutions. I then used parallel analysis, Very Simple Structure (VSS) criteria and the Velicer MAP criterion to compare solutions of varying complexity. The solution with one factor was selected and then estimated using CFA, and the fit of the model is assessed through the Normed Fit Index (NFI), Non-Normed Fit Index (NNFI), Comparative Fit Index (CFI), and root mean square error of approximation (RMSEA), all of which are frequently used to assess the global fit of latent variable models. This was followed by the estimation of a confirmatory maximum likelihood two-bifactor model, as explained in Chalmers (2012), which is better suited for educational standardized test data where items within a domain are more strongly associated to each other than to items of other domains. I also computed the M2 statistic, as well as fit indices comparing the fitted to the the null model using the root mean square error of approximation (RMSEA), the Standardized Root Mean Square Residual (SRMSR); the Tucker Lewis index (TLI); and again the CFI. Finally, Table 15 in Online Appendix presents the results of the estimation of the graded response model (GRD) on the mathematics test.

The following steps depart from the traditional validation framework:

Step 4: Apply the mini-test challenge. The mini-test challenge is based on ideas borrowed from Rosenbaum (1984, 428). Excluding item j, the sum score $R_{(-j),[i]} = \sum_{l=1}^{J} y_l$ for $l \neq j$ for each child i, also known as rest score, may be taken as a mini-test and can be used to empirically assess if other extraneous factors affect the probability of correctly answering an item. Assuming the sum score on the mini-test to be the best predictor for correctly answering item j, relative risks should show other variables are unrelated to this probability. I estimated relative risks using a log-linked binomial model for a set of sociodemographic covariates (preterm, gender, migration background and SES; for the estimation of SES see

Table 14 in Online Appendix). For the mini-test challenge applied to the mathematics test, the two polytomous items in the mathematics test were recoded. If at least one out of four options was correct, the item was scored as correct, ignoring the gradation difficulty implied by the question. The other items were dichotomous. Twenty models for each item were estimated, and the relative risks with 95% confidence intervals are presented. After adjusting for ability measured by the mini-test, relative risks of probabilities p_{x_l}/p_{x_k} for categories l, k of a categorical covariate (e.g. gender or social class), or p_{x+1}/p_x for a continuous covariate X (e.g. the PPVT-4 score), should equal one or be close to one, without showing patterns besides expected random fluctuation (see SAS (2018) for this conceptualization of relative risks). If the mathematics test is fair, valid and reliable, then the mini-test should be the best predictor of the probability of correctly answering an item in a test.

Step 5: Check that the quantitative structure assumption holds in the data. This steps goes beyond assuming psychological constructs have a quantitative structure. The hypothesis of quantifiability of an attribute can be verified by checking if a specific functional relation among the set of respondents (A) and the set of items (Q) holds (see Luce and Tukey (1964)). These conditions are transitivity, antisymmetry and strong convexity for an ordinal relation. There are six more conditions for additivity to shown that an interval functional relation holds: associativity, commutativity, monotonicity, solvability, positivity and the Archimedean condition (Heene (2013)). Karabatsos (2001), Karabatsos (2018) and Domingue (2014) have provided researchers with methods to empirically assess a stochastic version of the axioms of additive conjoint measurement (ACM). Domingue (2014) elaborated the connection among these conditions and the single and double cancellation axioms, which are empirically testable (although higher order cancellation cannot be checked under this framework, see Karabatsos (2018, 324)). When these conditions hold, the claim that the responses to a standardized test yield an interval scale are plausible; otherwise an ordinal scale must be used, but their assumptions should be check too (see Step 6). Violations of the quantitative structure assumption are expressed in the percentage of comparisons of adjacent 3×3 matrices that do not comply to the single and double cancellation axioms. Such checks have been used in some empirical applications, but the use of ACM is far from common within psychometrics (see Domingue (2014), Dimitrov (2016) for examples). The sole assumption of the method is that the data have been dichotomously coded and observed with error. These checks are computationally intensive, as the number of items in the test or sample size increase, but I performed them on the three standardized tests. ELFRA-2P and PPVT-4 consist of dichotomous items of spoken and not spoken words and correct/incorrect answers in each case. The mathematics test contains two polytomous items that were recoded as dichotomous, as for the mini-test challenge. **Step 6:** Consider using ordinal psychometric models and check their properties. NIRT models, such as the monotone homogeneity model (MHM), may be used as an exploratory tool to study response patterns and establish which properties of the scales constructed from items are present in the data (Sijtsma and Ark (2016)), even if ACM checks show violations of assumptions for the hypothesized quantitative structure. The paper proposes that the three basic psychometric properties of undimensionality, monotonicity and local independence should hold in the data before considering fitting a model for ordinal scales, but an advantage of Mokken scales is that such scales can be constructed from subsets of items that do conform to this assumptions. When such properties do not hold, ordinal scales cannot be used either. Mokken scales are shown for the mathematics test for illustration purposes.

-Unidimensionality claims that manifest responses to items are caused by one single attribute, construct or skill. Unidimensionality is assumed by the three standardized tests. Most IRT models assume unidimensionality too, and even though multidimensional IRT modes exist, their use is rare. No unique method to assess unidimensionality exists, but in the traditional validation framework the examination of factor loadings and eigenvalues generated by EFA is considered sufficient. The MHM assess dimensionality of a scale by examining the behavior of a family of scalability coefficients—the H_{jk}, H_j and H coefficients as defined in Sijtsma and Ark (2016, 145)—as the requirement to conform a scale of weak, medium or strong association is explored. The automatic item selection procedure (AISP) with the genetic algorithm has been shown to perform better in simulation studies to examine structure, though there are some limitations when seeking to discover a truly underlying structure, as explained in Straat, Van der Ark, and Sijtsma (2013) through a simulation study.

-Monotonicity: The assumption of monotonicity refers to the item step response function (ISRF). A monotonic ISRF refers to $\mathbb{P}(Y_j \geq y_j|\theta)$ being a non decreasing on the latent attribute θ for all j items. As the construct increases, the probability of correctly answering an item should be higher and likewise more difficult items should require higher values of ability. Number of violations of monotonicity assumption are presented.

-Local Independence: Conditioning on the attribute θ, items j, k are independent for all pairs $(j, k))$. The indices W_1 and W_3 present items flagged for local independence violations.

-Invariant Item Ordering: This property does not correspond to the MHM, but to the double homogeneity model. It states that all items are scored in the same order by all individuals responding to the test, at all levels of ability; it was assessed by the H^T coefficient.

-Guttman Errors: On the basis of the outlier score G_+, the distribution of children and number of Guttman errors is presented in a plot for the original J items and the subset of items conforming to a Mokken scale.

Step 7: Construct Mokken subscales. This step is based on the results for the complete item battery for the mathematics test. Mokken subscales are those that conform to the three properties assumed by IRT models. Only scalable items at $c = 0.3$ are chosen and items causing violations of monotonicity or local independence are excluded.

Step 8: Contrast modeling alternatives. This step illustrates changes in effects that may result from ignoring properties of the scale or measurement errors as shown in Steps 5 and 6. I compare estimates of mathematics ability based on, first, IRT ability estimates, second, the sum of correct responses, and third, the sum of correct responses in a Mokken scale. The comparison extends to the linear regression model and the CLM—using as link function the standard normal density, i.e., the Probit model—for the probabilities of being at different quartiles of the distribution of ability. Differences between the two modeling strategies are assessed descriptively. Given that variance in ordinal models cannot be decomposed as in linear models and that estimated parameters associated with each group of covariates will change with the inclusion or exclusion of additional covariates, model assessment was done by estimates of changes in probabilities, i.e., average marginal effects (AME). In light of potential unobserved heterogeneity (omission of some observed and unobserved variables) average marginal effects are preferable, because these can be compared across groups and models; and are more stable given alternative model specifications (see Mood (2010, 80) and Agresti and Tarantola (2018)).

Step 9 Check Measurement Invariance (MI) for groups of interest. In this case, I checked MI for preterm children, girls, children with migration background and children from low-SES families. The items composing a scale might function differently for different groups of individuals. If a scale does not equally evaluate individuals from different groups, then a test might be biased and its items should be examined and potentially excluded, if on closer examination, they are found not to be measurement invariant. Penfield and Camilli (2006) discussed nonparametric methods, describing the generalized Maentzel-Hazel (gMH) test as a statistic that does not require the estimation of an IRT model with its corresponding assumptions to hold. This statistic was used with continuity correction and P-value adjustment for multiple comparisons by the Holm method for a comparison of more than two groups. The raw score was used to match children from each group and a threshold or cut-off value score was selected to classify the corresponding items as displaying DIF (see Magis et al. (2010) for details of the computation of the statistic). No method for testing measurement invariance has been found superior to any other. Future work should consider alternative statistics.

Only complete cases were used. Sample sizes differed depending on the test being assessed. Multiple imputation was not applied, because it makes use of associations already present in the data, which the analysis presented here aims to empirically assess. With the exception of relative risk estimates, no inferential results are presented and regression models are shown only for illustrative purposes; different sample sizes do not affect the results. The sample sizes used are shown in Table 12 and Table 13 in the Online Appendix. Only the univariate estimates of relative risks are estimated taking into consideration sample design.

Finally, researchers have warned against the use of P-values to assess if some desired property is present in the data. As explained in Wasserstein, Schirm, and Lazar (2019), statistical findings and arguments based on quantitative information ought to rely on a coherent set of pieces of evidence, of which P-values are not a part of. Although the paper reports P-values, I do not discuss findings using the terminology of "statistical significance".

Software All analysis were performed in R v. 3.5.0, Rstudio environment. Main packages used in the analysis include oglmx version 3.0.0.0, mokken v. 2.8.11, psych v. 1.8.4., ConjointChecks v. 0.0.9, difR v. 5.0, lavaan v. 0.6-2 and mirt v. 1.30, among others.

Results

Step 1: Table 1 shows the frequency distribution of responses to mathematics test items and percentage of missing values. None of the items has low variance or a substantial percentage of missing values. At this stage, all items can be considered for the next analysis.

-Step 2: Polychoric correlation are shown in Figure 1. These are on the middle to low range, and there are even slightly negative correlations. Not all items vary in the same direction, as would be expected from a mathematics test, unless skill in some types of questions is negatively associated to others, which seems implausible. These items were however not excluded at this stage. Coefficients Omega hierarchical $\omega_h = 0.797$, Chronbach's $\alpha = 0.873621628446539$ and Guttman's $\lambda_6 = 0.764$ show high values; inter-item and drop-item statistics presented in

Table 1: Mathematics Test Items

	Missing		Incorrect or Wrong		Correct: Right answer | 1 out of 3		Correct: 2 out of 3		Correct: 3 out of 3	
	n	%	n	%	n	%	n	%	n	%
z17s	0	0.00	575	27.38	440	21.72	615	30.5	410	20.4
z021	0	0.00	948	45.94	1092	54.06	-	-	-	-
v181	0	0.00	1350	66.05	690	33.95	-	-	-	-
z161	3	0.15	649	31.24	1388	68.76	-	-	-	-
r14s	6	0.29	105	4.76	293	14.44	302	14.9	1334	65.9
d191	8	0.39	1040	51.10	992	48.90	-	-	-	-
z051	10	0.49	1383	68.00	647	32.00	-	-	-	-
g151	10	0.49	578	28.44	1452	71.56	-	-	-	-
r131	12	0.59	1369	67.35	659	32.65	-	-	-	-
g111	15	0.74	1749	86.36	276	13.64	-	-	-	-
z121	16	0.78	292	14.14	1732	85.86	-	-	-	-
v041	21	1.03	1358	67.05	661	32.95	-	-	-	-
z081	21	1.03	1893	93.73	126	6.27	-	-	-	-
d091	23	1.13	199	9.83	1818	90.17	-	-	-	-
z201	26	1.27	1236	61.33	778	38.67	-	-	-	-
g101	29	1.42	394	19.56	1617	80.44	-	-	-	-
z011	29	1.42	1374	68.20	637	31.80	-	-	-	-
r071	43	2.11	991	49.55	1006	50.45	-	-	-	-
d031	31	1.52	1422	70.71	587	29.29	-	-	-	-
v061	32	1.57	1138	56.67	870	43.33	-	-	-	-

Table 2 indicate sufficient internal reliability of the mathematics scale.

-Step 3: From the results of the exploratory factor analysis as shown in Table 3, the different criteria point to one factor as the best solution. As shown in Table 4 and Table 5, taken together, the items in the mathematics test have appropriate fit indices according to typical CFA and the bi-factor model criteria, substantiating the claim that items are measuring the unidimensional underlying construct of mathematics ability in children. The graded response model, for which some own calculations are presented in Table 15 and Figure 5 in the Online Appendix, and which is also presented in Petersen and Gerken (2018), shows evidence of appropriate fit, meaning that the estimates of mathematics ability in children are reliable and consistent, and presumably valid as well.

-Step 4: The cognitive construct of mathematics ability in children was subjected to the mini-test challenge (the short panel does not allow to assess the empirical correlations to later measures or outcomes). Figure 2 presents the relative risk associated with the mini-test for

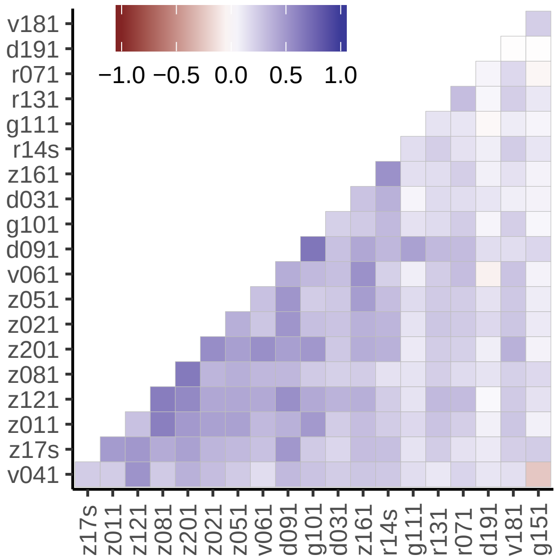

Figure 1: Polychoric Correlation for Mathematics Test Items

each item, ordered from easiest to hardest according to the percentage of correct responses. One extra point in the mini-test increases the probability of correctly answering the item excluded from the mini-test; these are, as expected, all greater than one. Moreover, the mini-test's predictive capacity increases with the difficulty of the items. This result validates the use of the test in the prediction of a correct response. Nevertheless, and as shown in Panels A, B, C and D of Figure 3, other covariates also remain predictive of a correct response in the items, even after controlling for mathematics ability as measured by the mini-test. For the language ability tests ELFRA-2P and PPVT-4 showed in panels E and F, no association with the probability of correctly answering an item is seen; but this is not the case for low-SES children. For 13 out of 20 items, the relative risks are below 1, meaning that there was a lower chance of correctly answering 13 items in the test despite controlling for mathematics ability as measured by the mini-test. The more difficult items also show effects for being born preterm and also for being a girl (some even in a positive direction). With the exception of SES, the effects of the other covariates do not follow a pattern, but it is possible that the total score captured something other than mathematics ability. Some noise might have permeated the measurement operation.

Table 2: Internal Reliability Coefficients for Mathematics Test Items

	Chronbach's Alpha	Guttman's Lambda 6	Average interitem cor.	Median interitem cor.	Cor. with score (corrected)	Drop item cor.
z17s	0.865	0.919	0.252	0.240	0.614	0.552
z021	0.863	0.925	0.250	0.238	0.654	0.597
v181	0.871	0.975	0.263	0.249	0.451	0.373
z161	0.865	0.975	0.253	0.238	0.605	0.542
r14s	0.868	0.948	0.256	0.244	0.549	0.479
d191	0.880	0.976	0.278	0.256	0.214	0.124
z051	0.864	0.928	0.251	0.238	0.635	0.575
g151	0.879	0.990	0.277	0.256	0.231	0.141
r131	0.871	0.953	0.262	0.249	0.468	0.392
g111	0.876	0.944	0.271	0.254	0.325	0.239
z121	0.860	0.913	0.245	0.238	0.728	0.680
v041	0.870	0.945	0.261	0.243	0.479	0.404
z081	0.864	0.914	0.251	0.243	0.634	0.574
d091	0.859	0.906	0.243	0.238	0.765	0.722
z201	0.860	0.926	0.244	0.238	0.750	0.705
g101	0.866	0.905	0.254	0.238	0.590	0.524
z011	0.863	0.903	0.250	0.238	0.655	0.597
r071	0.872	0.983	0.263	0.251	0.441	0.363
d031	0.870	0.971	0.261	0.246	0.475	0.399
v061	0.866	0.964	0.254	0.238	0.581	0.515

Table 3: Criteria for Number of Factors in Mathematics Test Items

Factors	VSS 1	VSS 2	MAP	Parallel FA
1	0.498	0.000	0.005	3.130
2	0.364	0.413	0.007	0.435
3	0.319	0.413	0.011	0.179
4	0.251	0.359	0.015	0.157
5	0.246	0.343	0.021	0.128
6	0.250	0.325	0.027	0.084

Step 5: The test for an ACM structure in the mathematics test responses reveal, as shown in Table 6, several violations of the single and double cancellation conditions. The weighted and unweighted proportion of violations are high when compared to the 2% for the unweighted and 1% for weighted violations for data simulated from a Rasch model and subjected to the same checks (Domingue (2014)). Heene (2013) has warned that such results would be obtained for many data sets that fit parametric IRT models, even when violations of its assumptions are present, so the result is not surprising. None of these standardized tests satisfies the conditions of an interval scale (i.e. that the the differences between the units on test scores are of an equal interval); they do not fulfill the assumptions of a quantitative attribute despite the evidence in favor of appropriate fit in accordance with the standards

Table 4: Confirmatory Factor Analysis for Mathematics Test Items

Chi-Square	d.f.	P-value	NFI	NNFI	CFI	RMSEA
322	170	1.64e-11	0.972	0.985	0.987	0.021

Table 5: Full-Information Item Bi-factor and Two-Tier Analysis for Mathematics Test Items

M2	d.f.	P-value	RMSEA	SRMSR	TLI	CFI
237	146	2.67e-06	0.018	0.027	0.984	0.988

cited above. However, these scales might still provide ordinal information.

-Step 6: Researchers might claim that measurement at an ordinal level, which is not as strict as an interval level in the parametric IRT models, is still possible. The results of the empirical assessment of unidimensionality, monotonicity and local independence assumptions as well as invariant item ordering are shown in Table 7 and Table 8. The AISP with the genetic algorithm shows that some of the mathematics items are unscalable. The progression of the mathematics scale by increasing the threshold c as shown in Table 7 reveals that 7 items are unscalable and 2 items belong to another scale. No items were found to have $H_j < 0$, but inter-item scalibility coefficients $H_{jk} < 0$ were present as well as monotonicity and local independence violations. A weak scale ($c = 0.3$) was formed by 11 items. These form an ordinal scale (Mokken scale; MKS) without either violations of monotonicity or of local independence. An improvement is observed in the psychometric properties of this subscale. Figure 4 presents the distribution of Guttmann errors for the full and the MKS scales; these contain a considerable number of Guttman errors, but the MKS shows fewer.

-Step 7 Figure 5 presents standardized coefficients and average marginal effects for the different analytic strategies described in the methods section. The parametric IRT (PIRT) and the sum score of the 20 items of the test contain forms of measurement error as shown in the previous sections. The estimated coefficients for the preterm and migration background covariates may have been diminished because of this. Using the Mokken scale as a linear scale showed larger effects for the selected covariates. However, when considering effect sizes, as

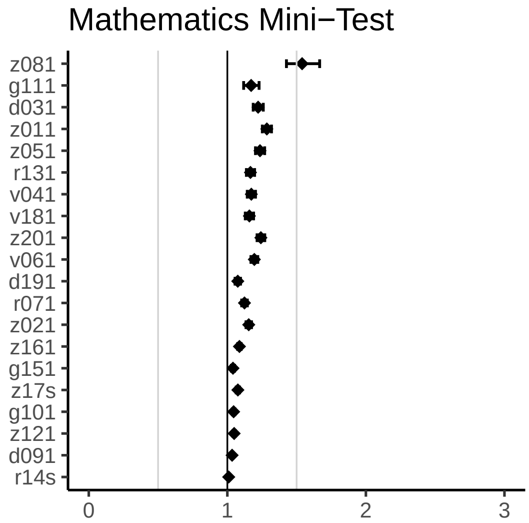

Figure 2: Mathematics Rest-score Risk Ratio in the Mini-Test Challenge

Table 6: ACM Checks for Mathematics Test, PPVT-4 and ELFRA-2P

	Adjacent 3x3 Matrices	
Instrument	Weighted Mean	Unweighted Mean
ELFRA-2 Productive subscale	18.1	40.8
PPVT-4	12.5	42.4
Mathematics Test	17.9	22.9

shown in Table 9, the differences between the linear regression and the CLM were noteworthy. AMEs, which were taken as overall effects of the predictor variables, showed that low-SES had the largest effects. By contrast, according to the linear model, language abilities had the largest main effect. Furthermore, the effects of being preterm or having migrant background were larger when using the Mokken scale when compared to the corresponding AME on the full scale; whereas for being from a low-SES background these effects were smaller when using the Mokken scale than the full scale. The direction of the effect of migration background changes comparing the PIRT to the Mokken scale. Noticeable though small differences were found in this comparative exercise, but the point is that different conclusions might be drawn from conceiving data differently, as argued in Liddell and Kruschke (2018).

Step 8 From the previous analysis, the data of the mathematics tests only warrants the use

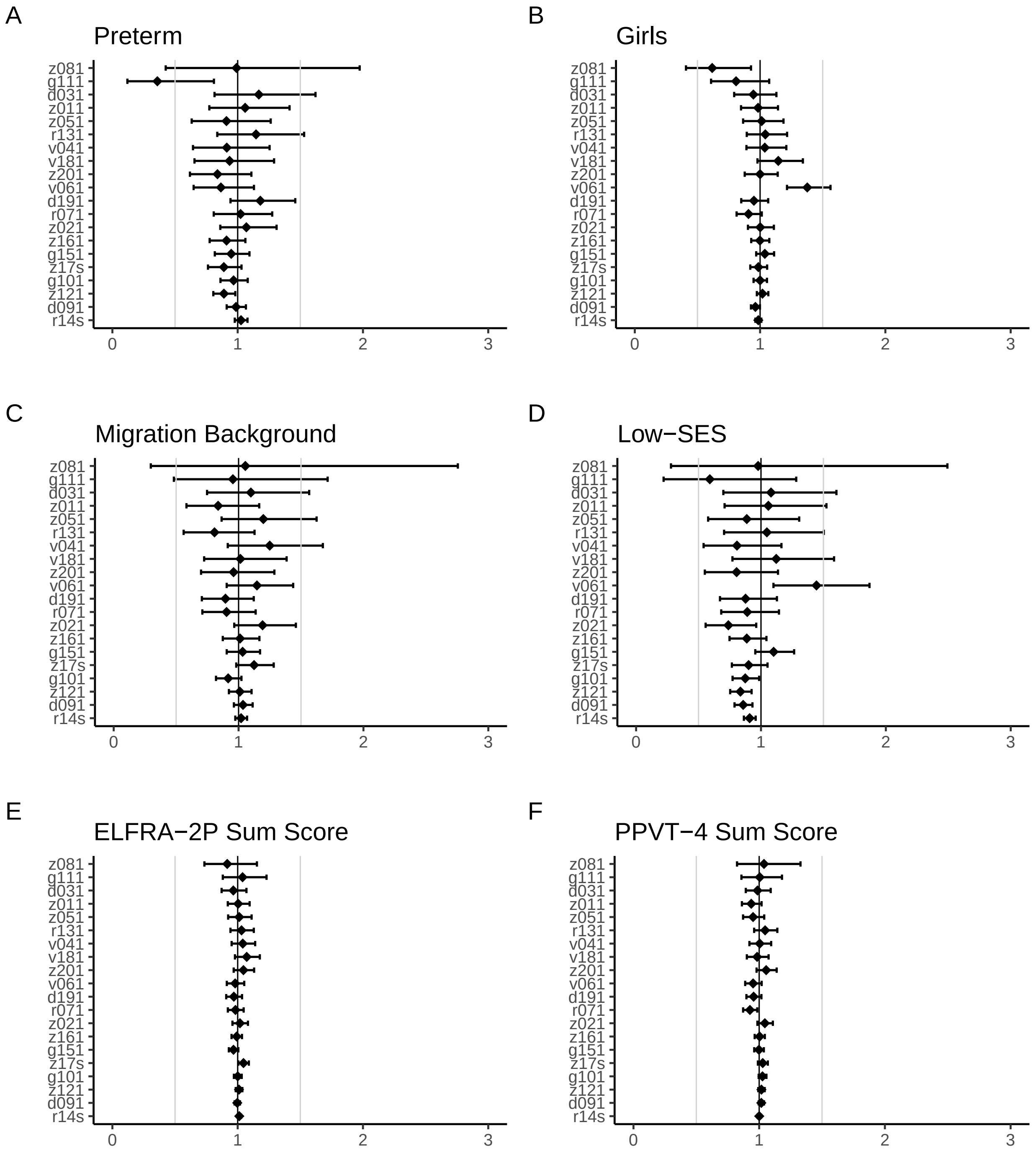

Figure 3: Relative risks for Various Covariates in the Mini-Test Challenge

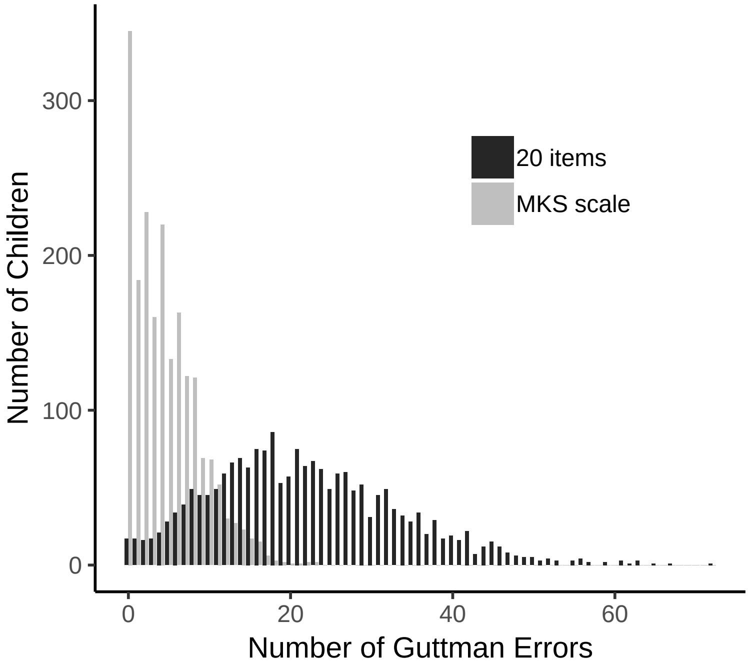

Figure 4: Number of Gutman Errors and Number of Children in the Mathematics Test Items and their Mokken Subscale

Table 7: AISP Genetic Algorithm for Mathematics Test Items

	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
z17s	1	1	1	1	1	1	1	0	1	1	0	0
z021	1	1	1	1	1	1	1	1	0	0	0	0
v181	1	1	1	1	1	0	0	2	0	0	0	0
z161	1	1	1	1	1	1	1	1	0	0	0	0
r14s	1	1	1	1	1	1	2	3	3	0	0	0
d191	2	2	0	0	0	0	0	0	0	0	0	0
z051	1	1	1	1	1	1	1	1	3	0	0	0
g151	1	1	0	2	0	0	0	0	0	0	0	0
r131	1	1	1	1	1	0	2	0	0	0	0	0
g111	1	1	1	1	0	0	0	0	0	0	0	0
z121	1	1	1	1	1	1	1	1	1	1	1	1
v041	2	2	1	1	1	1	0	3	0	0	0	0
z081	1	1	1	2	1	1	1	1	1	1	1	1
d091	1	1	1	1	1	1	1	1	1	1	1	2
z201	1	1	1	1	1	1	1	1	1	1	1	0
g101	1	1	1	1	1	1	1	2	2	0	0	0
z011	1	1	1	1	1	1	1	1	1	1	1	1
r071	1	1	1	1	0	0	0	0	0	0	0	0
d031	1	1	1	1	1	0	0	0	0	0	0	0
v061	1	1	1	1	1	1	1	0	2	0	0	2

of the CLM for the Mokken scale; this scale does fulfill the conditions of an ordinal scale. However, the Mokken scale of mathematics ability does not conform to the properties of a quantitative attribute either, given that the weighted average number of violations is 11.393 and the unweighted 17.07. Such treatment of an ordinal variable is not warranted by the properties of these subscale.

Step 9 The final step concerns DIF in the different standardized tests. Table 10 presents items that were flagged as DIF. Of special concern are those items for the mathematics test that show much higher odds of being correctly answered for groups different than the focal groups, as displayed in Table 11. Results suggest DIF in all three of these standardized tests. The PPVT-4, measuring verbal ability, has the largest number of items flagged as DIF. These high percentages might be indicators that the test is biased especially with regard to

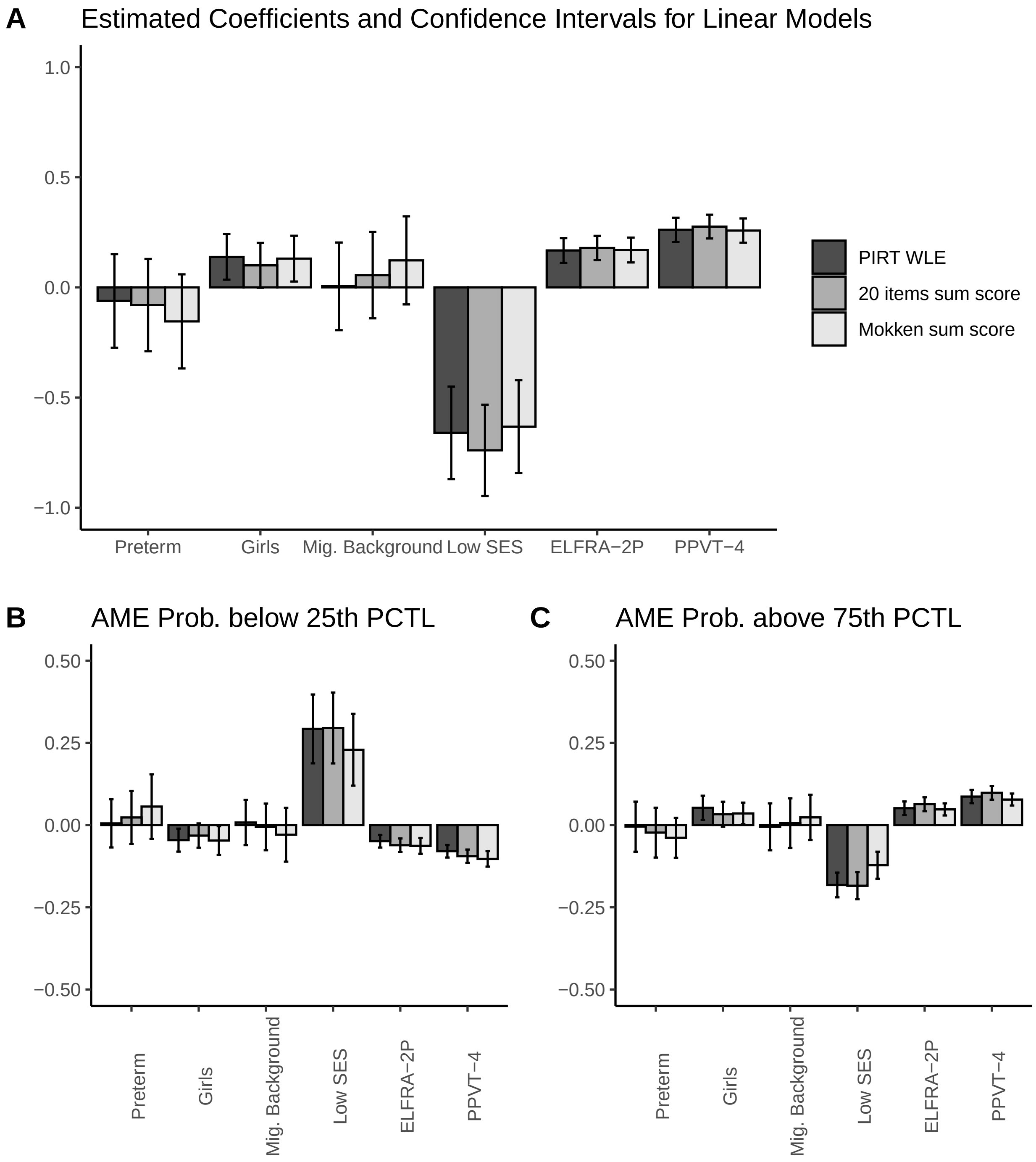

Figure 5: Standardized Estimated Coefficients in Linear Model (Panel A) and Average Marginal Effects from Cumulative Link Model (Panels B and C) for Scores on Mathematics Test

Table 8: Mokken Scale Analysis Results for Mathematics Test Items

	Mathematics Test	
Statistic	Complete: 20 items	Subscale: 11 items
Number of Unscalable Items at c=0.3	7	0
Number of Scales	2	1
Scalability Index H	0.236	0.378
Number of negative item-scale scalbility	0	0
Number of negative inter-item scalability	4	0
Monotonicity Violations	1	0
Number of Flagged Items W 1 Index	3	0
Number of Flagged Items W 3 Index	6	0
H_T	0.47	0.539

children with migration background, as well as children from low-SES families. The group of preterm children does not show notable differences, although some of the items are regarded as presenting DIF. More analysis is needed to confirm bias in these standardized cognitive tests, but the results already provide initial hints.

The results, partial as they may be, especially for the EFLRA-2P and the PPVT-4, which were mostly because of space limitation, show that these three standardized tests do not show the properties of an interval scale, nor do they seem unbiased for the groups here chosen. The items that fit the assumptions of IRT in the Mokken mathematics ability scale, excluding items flagged as DIF for each of the four covariates here examined (i.e., items z17s, v061, z121 and r14s), would leave the mathematics scale with only 7 items on which ordinal comparisons can be safely made.

The relative risk of being among the lowest scoring group (correctly answering 0, 1 or 2 out of 7 mathematics questions, which correspond to the 25th percentile of children's scores) in the Mokken scale of mathematics ability without including flagged DIF items is, for preterm children 1.476 times the risk than fullterm babies (C.I: [1.198, 1.798]); for girls 0.983 times the risk than boys (C.I: [0.886, 1.09]); for children with migration background 1.307 times the risk than non migrant background children (C.I: [1.124, 1.512]); and for children of parents in the

Table 9: Effect Size Estimates from Linear Models in Panel A

Dependent Variable	Covariates	SSR	d.f.	Eta	Delta
PIRT WLE	Preterm	0.26	1	0.0003	0.0002
	Gender	5.58	1	0.0058	0.0047
	Migration Background	0	1	1.71e-06	1.38e-06
	Socioeconomic Status	33.63	3	0.0337	0.0281
	ELFRA-2P	27.82	1	0.0280	0.0233
	PPVT-4	71.03	1	0.0686	0.0594
20 items Sum Score	Preterm	0.45	1	0.0005	0.0004
	Gender	2.91	1	0.0031	0.0024
	Migration Background	0.24	1	0.0003	0.0002
	Socioeconomic Status	42.59	3	0.0435	0.0356
	ELFRA-2P	31.61	1	0.0327	0.0264
	PPVT-4	79.13	1	0.0779	0.0662
Mokken Sum Score	Preterm	1.65	1	0.0017	0.0014
	Gender	4.97	1	0.0051	0.0042
	Migration Background	1.18	1	0.0012	0.0010
	Socioeconomic Status	32.72	3	0.0325	0.0274
	ELFRA-2P	28.45	1	0.0284	0.0238
	PPVT-4	69.12	1	0.0663	0.0578

least well off socioeconomic status (SES) 2.202 times the risk than children of the wealthiest, higher educated and better employed parents (C.I: [1.872, 2.59]). These are smaller, but still important social inequalities when compared to the ones discussed at the start of the paper. These inequalities are based on ordinal information and not on a metric “measure” of mathematical ability as presumed by the estimation of an IRT model. Such an ordinal conception of inequalities would lend itself to a different and more fruitful discussion of social inequalities in cognitive constructs.

Discussion

IRT aims at modeling the interaction between a person’s ability as a latent and unobserved trait and a given item stimulus (Borsboom (2006)), but does not guarantee that their underlying assumptions hold in any given scale formed from standardized test. Given that

Table 10: Number and Proportion of Items with DIF by Generalized Mantel-Haezel Test

	Preterm		Girls		Mig. Background		Low-SES	
	j	%	j	%	j	%	j	%
ELFRA-2 Productive subscale	14	5.38	67	25.8	92	35.4	78	30.00
PPVT-4	6	3.66	38	23.2	27	16.5	16	9.76
Mathematics Test	5	25.00	7	35.0	6	30.0	3	15.00

only IRT models are said to guarantee that the estimated abilities conform to an interval scale for comparison of groups, interest in assessing whether test scores from popular standardized tests conform to a quantitative structure is priority for social science research. A striking result from this empirical assessment is that none of the scales with the full battery of items displayed the less demanding properties of a Mokken scale, nor did they display those of an interval scale, because they violated the assumptions of a quantitative structure, which is the most fundamental one in psychometric models. Although the scales studied here have been validated in similar populations of children using the *Standards for Educational and Psychological Testing* as guidelines, from a more assumption-free point of view there is no support for their use as interval measures of ability in children. The results imply that a sum of correct responses—scaled with an IRT model or not, on the mathematics test or on the PPVT-4, or of words said by the child as reported by parents (ELFRA-2P)—should not be used to compare children by means of difference statistics. The use of test scores for comparisons on a difference scale must be warranted by the measurement operation and not by untested assumptions about what the attribute supposedly is (Velleman and Wilkinson (1993)).

Following the proposal by Dima (2018), the paper has presented a *comparative validation* scheme that may be applied to "measures" of underlying unobserved constructs. Building subscales with expected properties out of a battery of items that may violate underlying assumptions is motivated on "technical" grounds, but begs the question of whether a given pattern of responses is being caused by the construct being assessed, or whether it results

Table 11: Mathematics Items Flagged as DIF by Generalized Mantel-Haezel Test

Items	Preterm		Girls		Mig. Background		Low-SES	
	gMH	P-value	gMH	P-value	gMH	P-value	gMH	P-value
z17s	5.51	0.02	18.46	1.73e-05	5.96	0.01	5.96	0.01
z021	0.35	0.55	0.72	0.40	6.03	0.01	0.01	0.93
v181	0.85	0.36	7.28	6.97e-03	0.03	0.87	0.23	0.63
z161	0.00	0.96	0.00	0.95	4.92	0.03	0.57	0.45
r14s	14.78	1.21e-04	40.33	2.15e-10	65.12	6.66e-16	5.67	0.02
d191	5.35	0.02	2.14	0.14	1.21	0.27	1.85	0.17
z051	0.04	0.84	0.09	0.77	4.09	0.04	1.09	0.30
g151	0.07	0.79	0.14	0.71	2.07	0.15	3.28	0.07
r131	1.07	0.30	0.21	0.65	0.03	0.87	1.76	0.19
g111	3.87	0.05	6.18	0.01	3.13	0.08	1.31	0.25
z121	4.30	0.04	1.32	0.25	0.85	0.36	1.76	0.18
v041	0.31	0.58	0.20	0.65	0.03	0.86	1.00	0.32
z081	0.01	0.94	12.64	3.77e-04	0.81	0.37	0.12	0.73
d091	0.98	0.32	15.51	8.19e-05	0.05	0.82	0.26	0.61
z201	0.68	0.41	0.00	0.95	0.27	0.61	0.05	0.82
g101	1.51	0.22	0.02	0.89	2.41	0.12	3.74	0.05
z011	0.58	0.45	1.24	0.27	0.00	0.95	0.46	0.50
r071	0.59	0.44	0.97	0.32	0.04	0.83	0.19	0.67
d031	0.19	0.67	0.77	0.38	0.46	0.50	2.69	0.10
v061	0.34	0.56	28.27	1.05e-07	6.70	9.66e-03	13.67	2.18e-04

from the hypothesized and desired characteristics of a measurement instrument, as arguably happens when selecting items in PISA (Baird et al. (2017, 331)). The paper showed how the psychometric properties improved when validation was anchored in empirical tests of the properties that instruments purportedly measuring cognitive constructs should have. A reduction in measurement error was likely achieved at the expense of discarding information that did not fit fundamental assumptions of measurement models in psychology. This represents an alternative way in which measurement error may be understood and assessed without relying on the assumption of interval scales.

The results have direct consequences and legitimate educational sociologists' skepticism about interpreting social inequalities as a result of low abilities. Such inequalities might have partly resulted from biased and error laden measurement operations. From a pragmatist perspective a word of caution should be noted by researchers in educational inequality and social stratification. If hypothesized changes in the latent variable cannot be empirically traced to quantitative changes in the score produced by the standardized test, then information provided by test scores may make comparison across groups pointless, because is not possible to discern whether differences are caused by differences in the attribute, or in systematic or even random error, or maybe in both (Vautier et al. (2012)). The derivation of policy-relevant conclusions on the basis of a "measurement" instrument that is unable to distinguish the signal from the noise is problematic, especially when standardized tests might confuse one for the other. The paper does not speak against the use of testing (e.g. in a learning context, testing one-self or testing others might be efficient strategies to learn or to diagnose which skills have been mastered or in which areas understanding problems persists), but it does counsel against the use of test scores (at least the ones here studied) to compare children across time or social groups on difference statistics, which is, as this paper argues, an unjustified use.

Teachers' use of scores on standardized tests in classroom activities may reify social inequalities when differences of an ordinal type are misunderstood as being of an interval kind (see Dalziel (1998)). In a scenario in which child A has answered 12 out of 20 questions correctly, whereas child B got 6 out of 20, and child C only 3 out of 20, the difference in mathematics ability between B and C ($|b - c|$) compared to the difference between A and B ($|a - b|$) is not twice as much ($|a - b| \neq 2 \times |b - c|$, where $a, bandc$ are taken as measurement units of an unobserved construct, ability or competence. How far behind child C is from child B or A, or child B is from child A, remains fundamentally unknown and might even be a meaningless question to ask in the first place, yet is what first comes to mind when interpreting test scores as a reflection of underlying latent psychological constructs. The mini-test challenge cast serious doubts on the measurement of "mathematics ability" in children, when, for example, low-SES background negatively affects the chances of correctly answering a subset of items in a test, adjusting for "mathematics ability" as measured by the other items in the test. When teachers' judgement about who is more or less able are based partly on possibly invalid and biased standardized tests, the end result is a reification of the unequal social structures that produce unequal achievements as “measured” by test scores in the first place. DIF in these standardized tests casts doubts on the possibilities of straightforwardly comparing children. The school, as embedded in the reproduction of social inequalities, should be the focus of research on social inequalities and the role played by standardized tests in that mechanism fully worked out. If educational institutions, even from early on, incorporate similar standardized tests or standardized academic activities to evaluate children's abilities or achievements, social scientists should study in which ways the social might pervade the psychometric and focus on empirically assessing which properties a test has. Critics of standardized tests have repeatedly raised the issue of fairness in the assessment of cognitive constructs. Just as much as missing data problems have been widely accepted by the community of researchers, “missing fairness”" should be equally addressed. More specifically, this fairness problem should be addressed by studies on the emergence of social inequalities and especially in uses of test scores in educational settings to advance, promote (i.e. gate-keeping) or diagnose students' achievements, abilities or motivations, etc.

In light of these results, it is at least potentially misleading to claim that “skills beget skills” when measurement error notoriously affects the size of inequalities (not to mention the confounding implied by measurement bias in the estimation of effects from interventions in educational research in which the distribution of the error is unknown, as discussed in Kuroki and Pearl (2014)). Where early childhood inequalities in cognitive abilities exists, they might not have such pernicious repercussions in later stages of development, nor might they explain the persistence of inequality of educational opportunities despite targeted intervention programs in and out of school settings. Although not shown in the paper, interpretation of test scores may be hampered by a fundamental lack of validity. Empirical correlations to other outcomes, which are used to argue in favor of the validity of such standardized tests (using an outdated, highly criticized and severely misleading notion of *criterion validity*) might be the consequence of the test's reliability in capturing noise, and from educational institutions making extensive use of standardized tests in the classroom. Returning to the inequalities that I alluded to at the begging of the paper, the results of the analysis presented are worrisome, but for a different reason. The estimated social inequalities in mathematics ability, e.g., against children with migration background or low-SES children, are in large part the result of error prone assessments and possible socio-cultural biases in a standardized mathematics test, and not the reflection of inequalities embodied in children's brains, that are shown by in their (in)capacity to solve basic mathematics operations. The disconnection between brain development in a biological sense and what standardized tests are "measuring" could be the reason why targeted interventions, sometimes based on neuroscientific findings, appear to have no effect (e.g. Dillon et al. (2017)) or fade out over time.

From a social constructivist perspective, scores on standardized tests allegedly measuring psychological traits are a function of the social environments the child participates in; and a function of the measurement operation (Baird et al. (2017, 341)). Theoretically, scores on such tests relate to variables in a quantitative fashion, but researchers are still far from providing evidence that changes in the latent underlying construct relate or map to changes in the score produced by the instrument (see Michell (1997) and other works by this author on problems of measurement in psychology). Moreover, since assessment influences what is measured and how it is measured, more effort should be put in developing a framework for systematically evaluating DIF in its relationship to the extraneous factors that have been shown to affect "performance" of children. By assuming a specific underlying distribution for a trait, as in the "bell-curve", the "social" is inadvertently introduced into the psychometric and is deceivingly presented as truth. Researchers studying differences in tests scores should begin by establishing whether scales obtained from standardized tests conform to the assumptions of IRT models. When the empirical pattern of responses deviates from such assumptions, less demanding models to compare children should be used. Nonlinear methods for ordinal variables have considerable advantages over linear methods that assume that psychological constructs are quantifiable and measured on an interval scale. CLMs allow for more relevant questions about different conceptions of ability, as something that may be nonquantifiable, subjectively defined, changing, or not fixed, because it considers probabilities of events and ordering relationships among individuals.

It is important to bear in mind the symbolic violence exerted by the *dictum* of a test score (Bourdieu (1980) and Croizet (2011); and Désert, Préaux, and Jund (2009) for stereotype threat in children). Such effects are highly problematic given that concerns about equality of opportunities in education are at the center of sociological research on social inequality. Test-score-based arguments may end up presenting unfair assessments of children's abilities as inherent properties of the child that arise due to fundamental deficiencies embodied in their brains. The goods and services acquired by performing well on standardized tests in educational settings might be mistakenly taken to be a "reflection" of children's innate capacities, potential or efforts, sustaining false beliefs about what type of educational opportunities children deserve. Such testing instruments and their functioning within educational institutions might convert the benefits accrued unequally and unfairly among individuals into merit-based ones, further perpetuating the position of children from disadvantaged backgrounds as resulting from their own flaws (as in "intelligence" as "inherited", or a "lack of character" discourses that are still prevalent to this day (see Heckman, Humphries, and Kautz (2014), Heckman and Mosso (2014))). These are not metaphorical effects of measurement as understood in psychology and educational research but true consequences derived from potentially invalid interpretations of test scores. Instead then, it might be that "measurement error begets measurement error" or that a "biased assessment begets a biased judgement".

Social scientists have long documented the substantially lower academic chances of children from disadvantaged backgrounds (see Allmendinger, Ebner, and Nikolai (2018), Bourdieu and Passeron (1964)). Social stratification research suggests that data on competencies mediates the generation of such inequalities, which can be factored in by quantifying deficits in cognitive abilities. However, it is not clear whether inequalities in access to higher education or selection into the academic track—for example in the German education school system—result from such deficits in cognitive skills; or the deficits result from a long chain of standardized tests reliably measuring simultaneously signal and noise. Standardized tests and evaluations of children within the classroom might simultaneously create and certify the inequalities. Millet and Croizet (2016) showed how assessment affects children before entering primary school in kindergarten activities in France, and Grodsky, Warren, and Felts (2008) presented an overview of several issues regarding the use standardized tests within the US educational system.

This paper is hence an invitation to other social scientists interested in using test scores to draw inferences about social inequalities in educational achievement. Validation is a continuous process that should come at the start of analysis of standardized test data. The findings presented here lend credence to the replicability crisis in psychology (Loken and Gelman (2017)). The notion of having measured psychological constructs that can be replicated is ill-founded and psychologists might not have measured anything at all (Trendler (2013)), despite repeated claims to having done so. Even more so, the ontological basis of cognitive constructs (e.g. mathematics or language ability) might be more similar to a process happening in the brain than to a quantity residing presumably in the brain (Guyon (2018)), making the IRT enterprise useless. The lesson for applied researchers is to be mindful of problems of data on psychological standardized tests for assessing unobserved variables. It is too much to require of test developers that they empirically show how their test scores map onto measures of brain structures and functions, and yet, under a causal account of validity, this is a necessary requirement for establishing sound comparisons among children on a valid metric. Further requiring that observations obtained through these standardized tests conform to the conditions of a conjoint additive structure before scaling the data to produce an interval scale is too strict a criteria for psychological measures to meet, yet it is the only way one can speak of an interval metric, under a scientific conception of measurement in correspondence with the natural sciences.

To the social scientists, the message is that analysis of test and questionnaire data is considerably more complex than the *Standards for Educational and Psychological Testing* pretend them to be, but the problems associated with ignoring the three issues discussed at the begging of the paper are worrisome. For the psychological and psychometric researchers involved in developing tests and applying sophisticated mathematical models to scale standardized tests data: Here is your "atomic bomb" back!

Online Appendix

Descriptives for complete cases for each measurement instrument

Table 12: Sample Descriptive Statistics Complete Cases in NEPS SC1 5th Wave

				ELFRA-2 Productive at Wave 3				PPVT-4 at Wave 4				Mathematics Test at Wave 5			
Variable	Categories	N	%	Mean	Median	Min.	Max.	Mean	Median	Min.	Max.	Mean	Median	Min.	Max.
Preterm birth	Full-term	1122	93.81	152	161	2	260	48.4	54.0	0	116	0.054	0.048	-3.56	3.19
	Preterm	74	6.19	142	156	14	251	47.1	51.0	2	108	-0.067	-0.016	-3.56	3.19
Gender	Boy	565	47.24	144	155	4	260	48.7	55.0	0	112	-0.042	-0.010	-3.56	3.19
	Girl	631	52.76	158	167	2	260	48.0	53.0	0	116	0.125	0.108	-3.56	3.19
Migration Background	No Migration Background	1101	92.06	155	164	4	260	49.6	56.0	0	116	0.080	0.086	-3.56	3.19
	With migration background	95	7.94	112	121	2	260	33.5	34.0	2	77	-0.342	-0.303	-2.27	1.78
Social Class	SES Class 1	268	22.41	150	156	6	260	49.7	56.0	0	112	0.093	0.089	-2.35	3.19
	SES Class 2	539	45.07	162	172	6	260	51.6	59.0	0	116	0.229	0.229	-2.99	3.19
	SES Class 3	299	25.00	147	158	2	260	45.8	50.0	0	100	-0.088	-0.117	-3.33	2.06
	SES Class 4	90	7.53	109	102	7	260	33.6	35.5	0	83	-0.743	-0.935	-3.56	1.84
Total	-	1196	100.00	152	160	2	260	48.3	54.0	0	116	0.046	0.038	-3.56	3.19

Table 13: Sample Descriptive Statistics Complete Cases by Standardized Test in NEPS SC1

		ELFRA-2 Productive at Wave 3						PPVT-4 at Wave 4						Mathematics Test at Wave 5					
Variable	Categories	N	%	Mean	Median	Min.	Max.	N	%	Mean	Median	Min.	Max.	N	%	Mean	Median	Min.	Max.
Preterm birth	Full-term	1997	93.8	141.0	151	0	260	1580	94.1	46.5	52.0	0	121	1729	94.28	-0.001	0.012	-3.56	3.19
	Preterm	132	6.2	124.1	133	1	251	99	5.9	42.9	47.0	0	108	105	5.72	-0.196	-0.137	-3.56	3.19
Gender	Boy	1061	49.8	131.1	144	0	260	820	48.8	45.8	52.0	0	121	905	49.35	-0.092	-0.071	-3.56	3.19
	Girl	1068	50.2	148.8	158	1	260	859	51.2	46.8	52.0	0	116	929	50.65	0.066	0.094	-3.56	3.19
Migration Background	No Migration Background	1892	88.9	146.1	157	0	260	1503	89.5	48.1	54.0	0	121	1634	89.09	0.041	0.053	-3.56	3.19
	With migration background	237	11.1	91.0	82	0	260	176	10.5	30.9	31.5	0	86	200	10.90	-0.442	-0.390	-3.31	2.53
Social Class	SES Class 1	459	21.6	145.2	153	5	260	365	21.7	47.8	54.0	0	112	408	22.25	0.066	0.065	-2.66	3.19
	SES Class 2	851	40.0	156.8	167	2	260	700	41.7	51.0	58.0	0	117	759	41.38	0.209	0.207	-2.99	3.19
	SES Class 3	562	26.4	130.0	140	0	260	432	25.7	43.4	47.0	0	121	487	26.55	-0.152	-0.137	-3.33	2.53
	SES Class 4	257	12.1	96.9	88	1	260	182	10.8	32.3	34.0	0	98	180	9.81	-0.741	-0.863	-3.56	1.84
Total	-	2129	100.0	140.0	151	0	260	1679	100.0	46.3	52.0	0	121	1834	100.00	-0.012	0.002	-3.56	3.19

Latent Class Analysis for the Social Class Structure Latent class analysis (LCA) was used to identify homogeneous groups in the data using three widely used socioeconomic characteristics: a) the educational level of mother and father; b) the occupations of mother and father classified by the occupational class structure of Eriksson and Goldthorpe (see Evans (1992) for an overview of its underlying claims); and c) the monthly household adjusted income level as reported by the child's parents. These characteristics are associated to socioeconomic status (SES), an unobserved variable. An inductive or formative model is used to estimate SES, as in latent class analysis following the work of Savage et al. (2013). Although researchers tend to associate an order among the set of latent classes, these classes constitute different categories without order. An assumption in LCA is that the different variables making up the latent classes are assumed to be conditionally independent given that the observations, in this case children, belong to the same class. The number of groups or classes was chosen based on statistical criteria, given that no theoretical number of SES strata is acknowledged. The model with 4 latent classes was chosen as providing the best fit (results not shown). As seen in Table 14, classification in four groups was still interpretable. Low-SES (Class 4) are children whose parents are mostly low educated, have low occupational attainment and are in the low category of household income. The most relevant group of contrast is against children with highly educated parents, high ranking occupations and a high household income (class 2), which serves as the reference category in all analysis in the paper.

Table 14: Latent Class Analysis for Social Class Structure in NEPS SC1 Cohort at Wave 1

		Latent Classes			
Variable	Categories	Class 1	Class 2	Class 3	Class 4
Mother's educational level	No degree or vocational/voluntary degree (Haupt-, Real-, Volksschulabschluss)	2.32	0.00	2.79	51.52
	Technical/applied or Civil Servant	33.94	6.11	48.97	39.46
	Technical Degree (Fachhochschulreife)	47.57	10.30	34.73	8.86
	University Education	16.17	83.59	13.50	0.17
Father's educational level	No degree or vocational/voluntary degree (Haupt-, Real-, Volksschulabschluss	1.36	0.31	6.61	45.73
	Technical/applied or Civil Servant	16.51	8.91	71.24	46.52
	Technical Degree (Fachhochschulreife)	39.31	8.20	22.16	5.21
	University Education	42.82	82.58	0.00	2.55
Mother's EGP occupational class	I and II	44.40	93.15	48.64	9.45
	IIIa and IIIb	45.69	5.15	40.91	50.19
	IVa, IVb and IVc	3.59	1.43	1.64	2.31
	V and VI	2.84	0.27	6.04	9.43
	VIIa and VIIb	3.48	0.00	2.78	28.63
Father's EGP occupational class	I and II	78.04	95.48	23.20	9.65
	IIIa and IIIb	11.41	2.08	16.89	11.56
	IVa, IVb and IVc	7.29	1.90	4.64	4.94
	V and VI	0.00	0.41	36.95	23.86
	VIIa and VIIb	3.26	0.13	18.32	50.00
Household adjusted monthly income in 2012 EUR	(0,1160]	15.02	7.36	28.94	80.44
	(1160,1620]	24.80	13.69	44.61	17.12
	(1620,2019]	34.11	29.49	23.47	1.62
	(2190,16200]	26.07	49.46	2.99	0.82

Parametric Item Response Theory (PIRT): partial credit model for polytomous item responses The partial credit model for polytomous item responses was fitted to the mathematics test data. The model provides estimates of item location and category threshold parameters (Thorpe and Favia (2012)). The fit of the model show convergence problems, which are possibly caused by item r14s not fitting the assumptions of the model; the probability of answering 2 out of 3 correct options in this item is never above 50% for some levels of ability. A transformation of this item to a dichotomous one generates the same output as found by Petersen and Gerken (2018). This might be the reason why in Petersen and Gerken (2018) this variable was recoded as dichotomous before fitting the model that was used to compute the NEPS SC1 mathematics ability estimates.

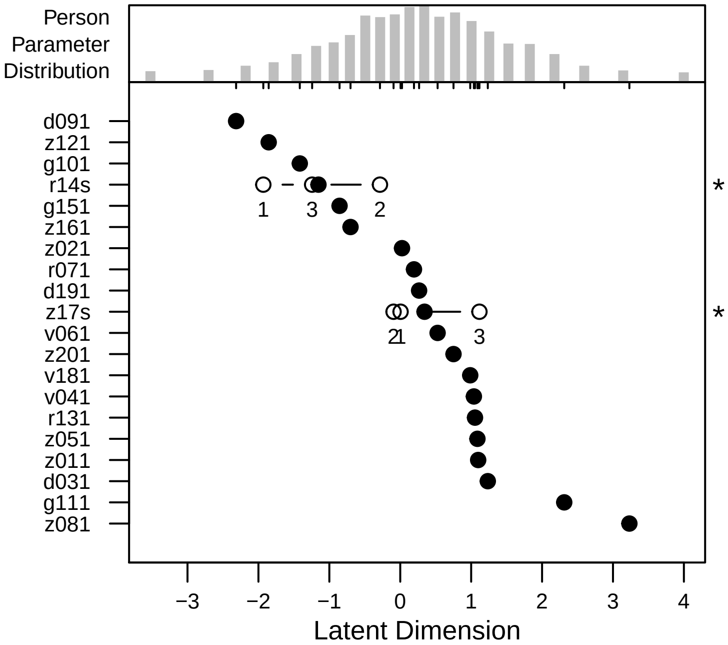

Figure 6: Person and Item Fit Plot for the Graded Response Model in Mathematics Test Items

Table 15: Graded Response Model Estimates for Mathematics Test Items

	Threshold: 1 vs. 0 | 1/3	Threshold: 2/3	Threshold: 3/3	Discrimination
z17s	-0.128	-0.44	1.1	0.71
z021	-0.153	-	-	1.42
v181	0.981	-	-	0.77
z161	-0.889	-	-	1.09
r14s	-2.865	-0.48	-2.6	0.56
d191	0.215	-	-	0.20
z051	0.821	-	-	1.16
g151	-3.386	-	-	0.28
r131	1.066	-	-	0.77
g111	4.369	-	-	0.44
z121	-1.455	-	-	1.90
v041	0.980	-	-	0.83
z081	2.089	-	-	1.91
d091	-1.937	-	-	1.57
z201	0.361	-	-	2.37
g101	-1.627	-	-	1.05
z011	0.675	-	-	1.69
r071	-0.028	-	-	0.67
d031	1.272	-	-	0.78
v061	0.299	-	-	1.15

Research Ethics

All parents of the children that participated in the NEPS SC1 gave consent on the collection of data, and their information is completely anonymized. The study conception was approved by an appropriate ethics committee.

References

Adams, Paul. 2006. "Exploring Social Constructivism: Theories and Practicalities." *Education* 34 (3). Taylor & Francis: 243–57. doi:10.1080/03004270600898893.

Agresti, Alan. 2012. *Analysis of Ordinal Categorical Data*. New Jersey: John Wiley & Sons.

Agresti, Alan, and Claudia Tarantola. 2018. "Simple Ways to Interpret Effects in Modeling

Ordinal Categorical Data." *Statistica Neerlandica* 72 (3). Wiley Online Library: 210–23. doi:10.1111/stan.12130.

Allmendinger, Jutta, Christian Ebner, and Rita Nikolai. 2018. "Soziologische Bildungsforschung." In *Handbuch Bildungsforschung*, edited by R. Tippelt and B. Schmidt-Hertha, 47–72. Springer. doi:10.1007/978-3-531-20002-6.

Association, American Educational Research, American Psychological Association, and National Council on Measurement in Education. 2014. American Psychological Association.

Baird, Jo-Anne, David Andrich, Therese N. Hopfenbeck, and Gordon Stobart. 2017. "Assessment and Learning: Fields Apart?" *Assessment in Education: Principles, Policy & Practice* 24 (3). Taylor & Francis: 317–50. doi:10.1080/0969594X.2017.1319337.

Ballou, Dale. 2009. "Test Scaling and Value-Added Measurement." *Education Finance and Policy* 4 (4). MIT Press: 351–83. doi:10.1162/edfp.2009.4.4.351.

Banerjee, Pallavi Amitava. 2016. "A Systematic Review of Factors Linked to Poor Academic Performance of Disadvantaged Students in Science and Maths in Schools." *Cogent Education* 3. Cogent OA: 17. doi:10.1080/2331186X.2016.1178441.

Batruch, Anatolia, Frédérique Autin, Fabienne Bataillard, and Fabrizio Butera. 2019. "School Selection and the Social Class Divide: How Tracking Contributes to the Reproduction of Inequalities." *Personality and Social Psychology Bulletin* 45 (3). SAGE Publications Sage CA: Los Angeles, CA: 477–90. doi:10.1177/0146167218791804.

Bauer, Andrea. 2014. "Methodenbericht. Neps Startkohorte 1-Haupterhebung 2016 B101." Bonn: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC1/5-0-0/SC1_MB5.pdf; infas Institut für angewandte Sozialwissenschaft GmbH.

Becker, Birgit. 2011. "Social Disparities in Children's Vocabulary in Early Childhood. Does Pre-School Education Help to Close the Gap?" *The British Journal of Sociology* 62 (1).

Wiley Online Library: 69–88. doi:10.1111/j.1468-4446.2010.01345.x.

Blossfeld, Hans-Peter, Nevena Kulic, and Moris Triventi. 2017. *Childcare, Early Education and Social Inequality: An International Perspective*. Edward Elgar Publishing.

Blossfeld, Hans-Peter, Hans-Günther Roßbach, and Jutta von Maurice. 2011. "Education as a Lifelong Process - the German National Educational Panel Study (Neps)." *Zeitschrift Für Erziehungswissenschaft* 14 (2): 19–34. doi:10.1007/s11618-011-0179-2.

Bond, Timothy N, and Kevin Lang. 2014. "The Sad Truth About Happiness Scales." 24853. *NBER Working Paper*. National Bureau of Economic Research. doi:10.3386/w24853.

Bond, Tymothy N., and Kevin Lang. 2013. "The Black-White Education Scaled Test-Score Gap in Grades K-7." *NBER Working Paper*, no. 19243. University of Wisconsin Press: 39. doi:10.3386/w19243.

Borsboom, Denny. 2006. "The Attack of the Psychometricians." *Psychometrika* 71 (3). Springer: 425–40. doi:10.1007/s11336-006-1447-6.

Borsboom, Denny, and Lisa D. Wijsen. 2017. "Psychology's Atomic Bomb." *Assessment in Education: Principles, Policy & Practice* 24 (3). Taylor & Francis: 440–46. doi:10.1080/0969594X.2017.1333084.

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111 (4). American Psychological Association: 1061–71. doi:10.1037/0033-295X.111.4.1061.

Bourdieu, Pierre. 1980. "Le Racisme de L'intelligence." In *Questions de Sociologie*, 264–68. Paris: Editions de Minuit.

Bourdieu, Pierre, and Jean-Claude Passeron. 1964. *Les Héritiers. Les étudiants et La Culture*. Paris: Editions de Minuit.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2013. "Measuring Test

Measurement Error: A General Approach." *Journal of Educational and Behavioral Statistics* 38 (6). Sage Publications Sage CA: Los Angeles, CA: 629–63. doi:10.3102/1076998613508584.

Breda, Thomas, Elyès Jouini, and Clotilde Napp. 2018. "Societal Inequalities Amplify Gender Gaps in Math." *Science* 359 (6381). American Association for the Advancement of Science: 1219–20.

Briggs, Derek C. 2013. "Measuring Growth with Vertical Scales." *Journal of Educational Measurement* 50 (2). Wiley Online Library: 204–26. doi:10.1111/jedm.12011.

Chalmers, R. Philip. 2012. "Mirt: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software* 48 (6): 1–29. doi:10.18637/jss.v048.i06.

Chan, Eric K.H. 2014. "Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments." In *Validity and Validation in Social, Behavioral, and Health Sciences*, edited by B.D. Zumbo, Eric K.H., and E.K.H. Chan, 9–24. Springer.

Croizet, Jean-Claude. 2011. "The Racism of Intelligence: How Mental Testing Practices Have Constituted an Institutionalized Form of Group Domination." In *The Oxford Handbook of African American Citizenship, 1865-Present*, edited by Bobo L.D., L. Crooms-Robinson, L. Darling-Hammond, M.C. Dawson, Gates H.J., G. Jaynes, and C. Steele. doi:10.1093/oxfordhb/9780195188059.013.0034.

Croizet, Jean-Claude, and Marion Dutrévis. 2004. "Socioeconomic Status and Intelligence: Why Test Scores Do Not Equal Merit." *Journal of Poverty* 8 (3). Taylor & Francis: 91–107. doi:10.1300/J134v08n03_05.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3). Wiley Online Library: 883–931. doi:10.3982/ECTA6551.

Dalziel, James. 1998. "Using Marks to Assess Student Performance, Some Problems and Alternatives." *Assessment & Evaluation in Higher Education* 23 (4). Taylor & Francis:

351–66. doi:10.1080/0260293980230403.

Dehaene-Lambertz, G., and E.S. Spelke. 2015. "The Infancy of the Human Brain." *Neuron* 88 (1). Elsevier: 93–109. doi:10.1016/j.neuron.2015.09.026.

Désert, Michel, Marie Préaux, and Robin Jund. 2009. "So Young and Already Victims of Stereotype Threat: Socio-Economic Status and Performance of 6 to 9 Years Old Children on Raven's Progressive Matrices." *European Journal of Psychology of Education* 24 (2). Springer: 207–18. doi:10.1007/BF03173012.

Dillon, Moira R., Harini Kannan, Joshuan T. Dean, Elizabeth S. Spelke, and Esther Duflo. 2017. "Cognitive Science in the Field: A Preschool Intervention Durably Enhances Intuitive but Not Formal Mathematics." *Science* 357 (6346). American Association for the Advancement of Science: 47–55. doi:10.1126/science.aal4724.

Dima, Alexandra L. 2018. "Scale Validation in Applied Health Research: Tutorial for a 6-Step R-Based Psychometrics Protocol." *Health Psychology and Behavioral Medicine* 6 (1). Taylor & Francis: 136–61. doi:10.1080/21642850.2018.1472602.

Dimitrov, Dimiter M. 2016. "An Approach to Scoring and Equating Tests with Binary Items: Piloting with Large-Scale Assessments." *Educational and Psychological Measurement* 76 (6). SAGE Publications Sage CA: Los Angeles, CA: 954–75. doi:10.1177/0013164416631100.

Domingue, Ben. 2014. "Evaluating the Equal-Interval Hypothesis with Test Score Scales." *Psychometrika* 79 (1). Springer: 1–19. doi:10.1007/s11336-013-9342-4.

Duncan, Greg J., Kathleen M. Ziol-Guest, and Ariel Kalil. 2010. "Early-Childhood Poverty and Adult Attainment, Behavior, and Health." *Child Development* 81 (1). Wiley Online Library: 306–25. doi:10.1111/j.1467-8624.2009.01396.x.

Elizabeth, S. Anderson. 2017. "What Is the Point of Equality?" In *Theories of Justice*, edited by A. Mancilla, 133–83. https://www.taylorfrancis.com/books/e/9781315236322/chapters/

10.4324/9781315236322-9; Routledge.

Ermisch, John. 2008. “Origins of Social Immobility and Inequality: Parenting and Early Child Development.” *National Institute Economic Review* 205 (1). SAGE Publications Sage UK: London, England: 62–71. doi:10.1177/0027950108096589.

Evans, Geoffrey. 1992. “Testing the Validity of the Goldthorpe Class Schema.” *European Sociological Review* 8 (3). Oxford University Press: 211–32. doi:10.1093/oxfordjournals.esr.a036638.

Farah, Martha J. 2017. “The Neuroscience of Socioeconomic Status: Correlates, Causes, and Consequences.” *Neuron* 96 (1). Elsevier: 56–71. doi:10.1016/j.neuron.2017.08.034.

Gamboa, Luis, and Erika Londoño. 2015. “Assessing Educational Unfair Inequalities at a Regional Level in Colombia.” *Lecturas de Economía*, no. 83. Universidad de Antioquia: 97–133. doi:10.17533/udea.le.n83a04.

Goussé, Marion, and Noémie Le Donné. 2014. “Why Do Inequalities in 15-Year-Old Cognitive Skills Increased so Much in France?” *SSRN*. Elsevier. doi:10.2139/ssrn.2533735.

Grimm, H., and H. Doil. 2006. “ELFRA 2-Elternfragebogen Für Zweijährige Kinder.” Göttingen: Hogrefe.

Grodsky, Eric, John Robert Warren, and Erika Felts. 2008. “Testing and Social Stratification in American Education.” *Annu. Rev. Sociol* 34. Annual Reviews: 385–404. doi:10.1146/annurev.soc.34.040507.134711.

Guyon, Herve. 2018. “Variables Latentes et Propriétés Mentales: Pour Une épistémologie Affirmée Pragmatiste et Réaliste.” *Psychologie Française*. Elsevier. doi:10.1016/j.psfr.2017.11.002.

Halle, Tamara, Nicole Forry, Elizabeth Hair, Kate Perper, Laura Wandner, Julia Wessel, and Jessica Vick. 2009. “Disparities in Early Learning and Development: Lessons from the Early Childhood Longitudinal Study-Birth Cohort (Ecls-B).” *Child Trends*. Washington, D.C.: https://www.childtrends.org/wp-content/uploads/2018/03/ECDisparities_ChildTrends_Jun2009.pdf, 39.

Heckman, James J. 2006. “Skill Formation and the Economics of Investing in Disadvantaged Children.” *Science* 312 (5782). American Association for the Advancement of Science: 1900–1902. doi:10.1126/science.1128898.

Heckman, James J., and Stefano Mosso. 2014. “The Economics of Human Development and Social Mobility.” *Annual Review of Economics* 6. Annual Reviews: 689–733. doi:10.1146/annurev-economics-080213-040753.

Heckman, James J., John Eric Humphries, and Tim Kautz, eds. 2014. *The Myth of Achievement Tests: The Ged and the Role of Character in American Life.* Chicago: University of Chicago Press.

Heene, Moritz. 2013. “Additive Conjoint Measurement and the Resistance Toward Falsifiability in Psychology.” *Quantitative Psychology and Measurement* 4 (246). Frontiers in Psychology. doi:10.3389/fpsyg.2013.00246.

Humphry, Stephen M. 2013. “A Middle Path Between Abandoning Measurement and Measurement Theory.” *Theory & Psychology* 23 (6). Sage Publications Sage UK: London, England: 770–85. doi:10.1177/0959354313499638.

Jednoróg, Katarzyna, Irene Altarelli, Karla Monzalvo, Joel Fluss, Jessica Dubois, Catherine Billard, Ghislaine Dehaene-Lambertz, and Franck Ramus. 2012. “The Influence of Socioeconomic Status on Children’s Brain Structure.” *PloS One* 7 (8). Public Library of Science: 9. doi:10.1371/journal.pone.0042486.

Johnson, H.M. 1936. *Pseudo-Mathetmatics in the Mental and Social Sciences.* Vol. 48. http://psychology.okstate.edu/faculty/jgrice/psyc3214/Johnson_1936_PseudoMath.pdf; The American Journal of Psychology.

Johnson, Sara B., Jenna L. Riis, and Kimberly G. Noble. 2016. “State of the Art Review:

Poverty and the Developing Brain." *Pediatrics* 137 (4). American Academy of Pediatrics: 18. doi:peds.2015-3075.

Kalil, Ariel, Rebecca Ryan, and Michael Corey. 2012. "Diverging Destinies: Maternal Education and the Developmental Gradient in Time with Children." *Demography* 49 (4). Springer: 1361–83. doi:10.1007/s13524-012-0129-5.

Karabatsos, George. 2001. "The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory." *Journal of Applied Measurement* 2 (4). https://pdfs.semanticscholar.org/8547/e64fd8d4f9efc7abd70a96496e0891eb21e3.pdf: 389–423.

———. 2018. "On Bayesian Testing of Additive Conjoint Measurement Axioms Using Synthetic Likelihood." *Psychometrika* 83 (2). Springer: 321–32. doi:10.1007%2Fs11336-017-9581-x.

Kuroki, Manabu, and Judea Pearl. 2014. "Measurement Bias and Effect Restoration in Causal Inference." *Biometrika* 101 (2). Oxford University Press: 423–37. doi:10.1093/biomet/ast066.

Kyngdon, Andrew. 2010. "Plausible Measurement Analogies to Some Psychometric Models of Test Performance." *British Journal of Mathematical and Statistical Psychology* 64 (3). Wiley Online Library: 478–97. doi:10.1348/2044-8317.002004.

Lacot, Emile, Mohammad H. Afzali, and Stéphane Vautier. 2016. "Test Validation Without Measurement." *European Journal of Psychological Assessment* 32. Hogrefe Publishing. doi:10.1027/1015-5759/a000253.

Lee, Dohoon, and Margot Jackson. 2017. "The Simultaneous Effects of Socioeconomic Disadvantage and Child Health on Children's Cognitive Development." *Demography* 54 (5). Springer: 1845–71. doi:10.1007/s13524-017-0605-z.

Liddell, Torrin M., and John K. Kruschke. 2018. "Analyzing Ordinal Data with Metric Models: What Could Possibly Go Wrong?" *Journal of Experimental Social Psychology* 79.

Elsevier: 328–48. doi:10.1016/j.jesp.2018.08.009.

Loken, Eric, and Andrew Gelman. 2017. "Measurement Error and the Replication Crisis." *Science* 355 (6325). American Association for the Advancement of Science: 584–85. doi:10.1126/science.aal3618.

Luce, R. Duncan, and John W. Tukey. 1964. "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology* 1 (1). Elsevier: 1–27. doi:10.1016/0022-2496(64)90015-X.

Magis, David, Sébastian Béland, Francis Tuerlinckx, and Paul De Boeck. 2010. "A General Framework and an R Package for the Detection of Dichotomous Differential Item Functioning." *Behavior Research MMethods* 42 (3). Springer: 847–62. doi:10.3758/BRM.42.3.847.

Mari, Luca, Andrew Maul, David Torres Irribarra, and Mark Wilson. 2017. "Quantities, Quantification, and the Necessary and Sufficient Conditions for Measurement." *Measurement* 100. Elsevier: 115–21. doi:10.1016/j.measurement.2016.12.050.

Maul, Andrew. 2017. "Rethinking Traditional Methods of Survey Validation." *Measurement: Interdisciplinary Research and Perspectives* 15 (2). Taylor & Francis: 51–69. doi:10.1080/15366367.2017.1348108.

Maul, Andrew, David Torres Irribarra, and Mark Wilson. 2016. "On the Philosophical Foundations of Psychological Measurement." *Measurement* 79. Elsevier: 311–20. doi:10.1016/j.measurement.2015.11.001.

Michell, Joel. 1997. "Quantitative Science and the Definition of Measurement in Psychology." *British Journal of Psychology* 88. Wiley Online Library: 355–83. doi:10.1111/j.2044-8295.1997.tb02641.x.

———. 2008. "Is Psychometrics Pathological Science?" *Measurement* 6 (1-2). Taylor &

Francis: 7–24. doi:10.1080/15366360802035489.

———. 2012. "Alfred Binet and the Concept of Heterogeneous Orders." *Quantitative Psychology and Measurement* 3 (261). Frontiers in Psychology: 8. doi:10.3389/fpsyg.2012.00261.

Millet, Mathias, and Jean-Claude Croizet. 2016. *L'école Des Incapables? La Maternelle, Un Apprentissage de La Domination.* Paris: La Dispute.

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It." *European Sociological Review* 26 (1). Oxford University Press: 67–82. doi:10.1093/esr/jcp006.

Nash, Roy. 2001. "Class,'Ability' and Attainment: A Problem for the Sociology of Education." *British Journal of Sociology of Education* 22 (2). Taylor & Francis: 189–202. doi:10.1080/01425690120054821.

Noble, Kimberly G., Laura E. Engelhardt, Natalie H. Brito, Luke J. Mack, Elizabeth J. Nail, Jyoti Angal, Rachel Barr, William P. Fifer, and Amy J. Elliott. 2015. "Socioeconomic Disparities in Neurocognitive Development in the First Two Years of Life." *Developmental Psychobiology* 57 (5). Wiley Online Library: 535–51. doi:10.1002/dev.21303.

O'Brien, Robert M. 1985. "The Relationship Between Ordinal Measures and Their Underlying Values: Why All the Disagreement?" *Quality & Quantity* 19 (3). Springer: 265–77. doi:10.1007%2FBF00170998?LI=true.

Penfield, Randall D., and Gregory Camilli. 2006. "Differential Item Functioning and Item Bias." In *Handbook of Statistics*, 26:125–67. Elsevier. doi:10.1016/S0169-7161(06)26005-X.

Petersen, Lara Aylin, and Anna-Lena Gerken. 2018. "NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 1 or Four-Year Old Children." *NEPS Survey Papers*, no. 45. https://www.neps-data.de/Portals/0/Survey%20Papers/SP_XLV.pdf.

Protopapas, Athanassios, Rauno Parrila, and Panagiotis G. Simos. 2016. "In Search of

Matthew Effects in Reading." *Journal of Learning Disabilities* 49 (5). SAGE Publications Sage CA: Los Angeles, CA: 499–514. doi:10.1177/0022219414559974.

Revelle, William, and Richard E. Zinbarg. 2009. "Coefficients Alpha, Beta, Omega, and the Glb: Comments on Sijtsma." *Psychometrika* 74 (1). Springer: 145–54. doi:10.1007/S11336-008-9102-Z.

Rosenbaum, Paul R. 1984. "Testing the Conditional Independence and Monotonicity Assumptions of Item Response Theory." *Psychometrika* 49 (3). Springer: 425–35. doi:10.1007/BF02306030.

Roßbach, Hans-Günther, W. Tietze, and Sabine Weinert. 2005. *Peabody Picture Vocabulary Test-Revised. Deutsche Forschungsversion Des Tests von Lm Dunn & Lm Dunn von 1981*. Bamberg/Berlin: Pearson.

Sachse, S., and W. v Suchodoletz. 2007. "Variabilität Expressiver Sprachleistungen Bei Zweijährigen Kindern Erfasst Mit Dem Elfra-2." *Sprache Stimme Gehör* 31 (3). Georg Thieme Verlag KG Stuttgart New York: 118–25. doi:10.1055/s-2007-982528.

Sachse, S., and W. Von Suchodoletz. 2007. "Diagnostische Zuverlässigkeit Einer Kurzversion Des Elternfragebogens Elfra-2 Zur Früherkennung von Sprachentwicklungsverzögerungen." *Klinische Pädiatrie* 219. Georg Thieme Verlag KG Stuttgart New York: 76–81. doi:10.1055/s-2006-942174.

SAS. 2018. "Usage Note 23003: Estimating a Relative Risk (Also Called Risk Ratio, Prevalence Ratio)." https://support.sas.com/kb/23/003.html.

Savage, Mike, Fiona Devine, Niall Cunningham, Mark Taylor, Li Yaojun, Johs Hjellbrekke, Brigitte. Le Le Roux, Sam Friedman, and Andrew Miles. 2013. "A New Model of Social Class? Findings from the Bbc's Great British Class Survey Experiment." *Sociology* 47 (2). Sage Publications Sage UK: London, England: 219–50. doi:10.1177/0038038513481128.

Schoon, Ingrid. 2010. "Childhood Cognitive Ability and Adult Academic Attainment:

Evidence from Three British Cohort Studies." *Longitudinal and Life Course Studies: International Journal* 1 (3): 241–158. doi:10.14301/llcs.v1i3.93.

Shear, Benjamin R., and Bruno D. Zumbo. 2014. "What Counts as Evidence: A Review of Validity Studies in Educational and Psychological Measurement." In *Validity and Validation in Social, Behavioral, and Health Sciences*, edited by B.D. Zumbo, Eric K.H., and E.K.H. Chan, 91–111. Springer. doi:10.1007/978-3-319-07794-9_6.

Sijtsma, Klaas, and L. Andries van der Ark. 2016. "A Tutorial on How to Do a Mokken Scale Analysis on Your Test and Questionnaire Data." *British Journal of Mathematical and Statistical Psychology* 70 (1). Wiley Online Library: 137–58. doi:10.1111/bmsp.12078.

Sijtsma, Klaas, and Rob R. Meijer. 2006. "Nonparametric Item Response Theory and Related Topics." In *Handbook of Statistics*, 26:719–46. doi:10.1016/S0169-7161(06)26022-X.

Solga, Heike, and Rosine Dombrowski. 2009. "Soziale Ungleichheiten in Schulischer Und Außerschulischer Bildung: Stand Der Forschung Und Forschungsbedarf." *Arbeitspapier, Bildung Und Qualifizierung.* Hans Boeckler Stiftung; https://www.boeckler.de/pdf/p_arbp_171.pdf.

Straat, J. Hendrik, L. Andries Van der Ark, and Klaas Sijtsma. 2013. "Comparing Optimization Algorithms for Item Selection in Mokken Scale Analysis." *Journal of Classification* 30 (1). Springer: 75–99. doi:10.1007/s00357-013-9122-y.

Tella, Patricia, Luciane da Rosa Piccolo, Mayra Lemus Rangel, Luis Augusto Rohde, Guilherme Vanoni Polanczyk, Euripides Constantino Miguel, Sandra Josefina Ferraz Ellero Grisi, Bacy Fleitlich-Bilyk, and Alexandre Archanjo Ferraro. 2018. "Socioeconomic Diversities and Infant Development at 6 to 9 Months in a Poverty Area of São Paulo, Brazil." *Trends in Psychiatry and Psychotherapy* 40 (3). SciELO: 232–40. doi:10.1590/2237-6089-2017-0008.

Thorpe, Geoffrey L., and Andrej Favia. 2012. "Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications." *Psychology Faculty*

Scholarship, no. 20. https://digitalcommons.library.umaine.edu/cgi/viewcontent.cgi?article=1019&context=psy_facpub.

Trendler, Günter. 2009. "Measurement Theory, Psychology and the Revolution That Cannot Happen." *Theory & Psychology* 19 (5). Sage Publications Sage UK: London, England: 579–99. doi:10.1177/0959354309341926.

———. 2013. "Measurement in Psychology: A Case of Ignoramus et Ignorabimus? A Rejoinder." *Theory & Psychology* 23 (5). Sage Publications Sage UK: London, England: 591–615. doi:10.1177/0959354313490451.

———. 2018. "Conjoint Measurement Undone." *Theory & Psychology* 29 (1). SAGE Publications Sage UK: London, England: 100–128. doi:10.1177/0959354318788729.

Van Laar, Colette, and Jim Sidanius. 2001. "Social Status and the Academic Achievement Gap: A Social Dominance Perspective." *Social Psychology of Education* 4 (3-4). Springer: 235–58. doi:10.1023/A:1011302418327.

Vautier, Stéphane, Michiel Veldhuis, Émile Lacot, and Nadine Matton. 2012. "The Ambiguous Utility of Psychometrics for the Interpretative Foundation of Socially Relevant Avatars." *Theory & Psychology* 22 (6). Sage Publications Sage UK: London, England: 810–22. doi:10.1177/0959354312450093.

Velleman, Paul F., and Leland Wilkinson. 1993. "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading." *The American Statistician* 47 (1). Taylor & Francis Group: 65–72. doi:10.1080/00031305.1993.10475938.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World Beyond 'P< 0.05'." *The American Statistician*. Taylor & Francis. doi:10.1080/00031305.2019.1583913.

Wehby, George L., and Antonio J. Trujillo. 2017. "Differences in Early Cognitive and Receptive-Expressive Neurodevelopment by Ancestry and Underlying Pathways in Brazil and Argentina." *Infant Behavior and Development* 46. Elsevier: 100–114. doi:10.1016/j.infbeh.2016.12.001.

Weinert, Sabine, Anja Linberg, Manja Attig, Jan-David Freund, and Tobias Linberg. 2016. "Analyzing Early Child Development, Influential Conditions, and Future Impacts: Prospects of a German Newborn Cohort Study." *International Journal of Child Care and Education Policy* 10 (7). Springer: 20. doi:10.1186/s40723-016-0022-6.

Winkelmann, Rainer, and Stefan Boes. 2006. *Analysis of Microdata*. Zurich: Springer Science & Business Media.

Wood, Robert. 1978. "Fitting the Rasch Model-a Heady Tale." *British Journal of Mathematical and Statistical Psychology* 31 (1). Wiley Online Library: 27–32. doi:10.1111/j.2044-8317.1978.tb00569.x.

Wright, Benjamin D. 1992. "Raw Scores Are Not Linear Measures: Rasch Vs. Classical Test Theory Ctt Comparison." *Rasch Measurement Transactions* 6 (1). https://www.rasch.org/rmt/rmt61n.htm: 208.

Zinn, Sabine, Ariane Würbach, Hans Walter Steinhauer, and Angelina Hammon. 2018. "Attrition and Selectivity of the Neps Starting Cohorts: An Overview of the Past 8 Years." *NEPS Survey Papers*, no. 34. Bamberg, Germany: https://www.researchgate.net/profile/Sabine_Zinn/publication/324727198_Attrition_and_Selectivity_of_the_NEPS_Starting_Cohorts_An_Overview_of_the_Past_8_Years/links/5adf291daca272fdaf8935ce/Attrition-and-Selectivity-of-the-NEPS-Starting-Cohorts-An-Overview-of-the-Past-8-Years.pdf; Leibniz Institute for Educational Trajectories.