2016 ■ Vol. 12 ■ no. 1

DOI: 10.20982/tqmp.12.1.p009

A unified factor-analytic approach to the detection of item and test bias: Illustration with the effect of providing calculators to students with dyscalculia

Minji K. Lee[a]✉, James J. Lee[b], Craig S. Wells[c] & Stephen G. Sireci[c]

[a]University of Massachusetts Amherst and Mayo Clinic
[b]University of Minnesota Twin Cities
[c]University of Massachusetts Amherst

Abstract ■ An absence of measurement bias against distinct groups is a prerequisite for the use of a given psychological instrument in scientific research or high-stakes assessment. Factor analysis is the framework explicitly adopted for the identification of such bias when the instrument consists of a multi-test battery, whereas item response theory is employed when the focus narrows to a single test composed of discrete items. Item response theory can be treated as a mild nonlinearization of the standard factor model, and thus the essential unity of bias detection at the two levels merits greater recognition. Here we illustrate the benefits of a unified approach with a real-data example, which comes from a statewide test of mathematics achievement where examinees diagnosed with dyscalculia were accommodated with calculators. We found that items that can be solved by explicit arithmetical computation became easier for the accommodated examinees, but the quantitative magnitude of this differential item functioning (measurement bias) was small.

Keywords ■ factor analysis, item response theory, measurement invariance, differential item functioning, dyscalculia

✉ lee.minji@mayo.edu

Introduction

A difference between populations in average test performance cannot be safely interpreted unless the test provides nearly unbiased measurements of the same attribute in both populations. A battery of tests satisfying this property to a good approximation is said to exhibit *measurement invariance* (Meredith, 1993; Millsap, 2007), and factor analysis has been used to test the adequacy of this approximation in many applications. When the analysis focuses on the individual items within a single test, rather than several such tests within a larger battery, the corresponding property of unbiased measurement is called absence of *differential item functioning* (DIF). Many of the methods for detecting DIF fit within the overall framework of item response theory (IRT) (Holland & Wainer, 1993; Penfield & Camilli, 2007).

IRT is a mild nonlinearization of standard factor analysis, and indeed this connection has been recognized since the initial formulation of IRT (Lord & Novick, 1968; Muthén & Lehman, 1985; Takane & de Leeuw, 1987; McDonald, 1999; Kamata & Bauer, 2008; Bartholomew, Knott, & Moustaki, 2011). The mainstream of IRT, however, has tended to develop without reference to this underlying unity. As a result it is possible to find entire textbooks devoted to specialized some aspect of IRT with scarce acknowledgement of earlier and entirely parallel developments in the linear factor-analytic models implicitly underlying classical test theory. Conversely, despite the efforts of many methodologists, the power and elegant insights of IRT have influenced mainstream psychological science very unevenly. This mutual neglect has the potential to do harm if it hinders the migration of well-justified procedures from one level of analysis to the other.

In this article we provide an exposition of bias detection and highlight the benefits of the unified framework.[1]

[1]The term *bias* has accrued a number of meanings in various literatures. Here we use the term in a sense that should be acceptable to most psychologists: a group difference in the conditional expectation of performance, even when the comparison is restricted to individuals of the same "latent ability" or "factor level" (e.g., Lubke, Dolan, Kelderman, & Mellenbergh, 2003). This sense of bias is often called *measurement bias*, as distinct from *prediction bias*; the latter refers to a group difference in the conditional expectation of an external criterion when a test is used for purposes of prediction (Millsap, 2007). Furthermore, in the education literature, what we call "measurement bias" is often labeled by the more neutral term "differential functioning," until evidence of prediction bias or non-psychometric considerations are felt to justify the use of the stronger term. To repeat, for purposes of this article, we feel that using the terms "differential functioning" and "measurement bias" interchangeably and neglecting the issue of prediction bias altogether are reasonable compromises.

Our presentation will tend to take linear factor analysis and measurement invariance at the test level as starting points, but the treatment of both test and item levels will be self-contained. As a result the article can be read as an introduction to IRT itself (cf. Waller, Tellegen, McDonald, & Lykken, 1996). Although our intended audience consists of psychologists who routinely use factor analysis in their research, we believe that our approach should be illuminating to practitioners working within either one of these two psychometric traditions; on each side of the divide, we highlight insights and practices that are applicable to the analysis of datasets at any level.

Our real-data example comes from a statewide test of mathematics achievement. Certain types of mathematics items are thought to impose an unfair burden on students diagnosed with *dyscalculia*—a condition characterized by inordinate difficulties with learning simple numerical concepts, attaining an intuitive grasp of numbers, and applying routine arithmetical procedures (Butterworth, 2010). Such students were allowed to use calculators during the administration of the test. The *Standards for Educational and Psychological Testing* emphasize that issues of validity are critical whenever accommodations are offered (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). It seems especially important to investigate the possibility of measurement bias in this particular case, where both the disability and its remedy seem to be intimately related to the measured trait itself.

The Linear Common Factor Model and Measurement Bias in Quantitative Tests

We can write the standard linear model of factor analysis as

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^p$ is a vector of test scores, $\boldsymbol{\mu} \in \mathbb{R}^p$ is a vector of intercepts, $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times q}$ is a matrix of factor loadings, $\boldsymbol{\theta} \in \mathbb{R}^q$ is a vector of scores on common factors, and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is a vector of regression residuals. We have deliberately departed from convention in the use of $\boldsymbol{\theta}$, the usual symbol for latent traits in IRT, to denote factor scores; our purpose is to emphasize the virtual identity of these two concepts. The first application of linear factor analysis was Spearman's original *g* model, in which various tests of mental ability are regarded as measurements of a single general factor.

Equation 1 has the falsifiable consequence that

$$\mathbb{E}\left(\mathbf{y}\mathbf{y}'\right) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$ is the covariance matrix of the common factors and $\boldsymbol{\Theta} \in \mathbb{R}^{p\times p}$ is the diagonal matrix of residual variances. The diagonality of $\boldsymbol{\Theta}$ should be regarded as a definition rather than an assumption (McDonald, 1981); by the *principle of local independence*, each pair of tests should be measuring uncorrelated noise in a subpopulation with a fixed value of $\boldsymbol{\theta}$ and hence no variation in the traits to be measured. The discrepancy between the covariance matrix implied by Equation 2 and the actual sample covariance matrix provides the basis for parameter estimation and testing the adequacy of a hypothesis restricting the dimensionality of the factor space (q), the pattern of nonzero loadings in $\boldsymbol{\Lambda}$, and the form of the factor covariance matrix $\boldsymbol{\Psi}$.

Distinct groups of examinees can exhibit distributions of observed test scores that differ in critical ways from the distribution characterizing a traditionally more privileged group. Here we adopt the terminology of *focal* and *reference* groups to label these two types of examinee populations. To obtain some assurance that the collection of tests measures the same traits and provides an unbiased estimate of relative standing regardless of group membership, we can determine whether the same factor model holds in both groups to a satisfactory approximation. With two or more groups, the implication of Equation 1 for the mean structure,

$$\mathbb{E}(\mathbf{y}) = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\kappa}, \quad \text{where } \mathbb{E}(\boldsymbol{\theta}) = \boldsymbol{\kappa}, \tag{3}$$

can be empirically tested by fixing one group's mean to the zero vector.

Table 1 outlines a nested sequence of increasingly comprehensive models testing measurement invariance. Notice that no model imposes the restriction $\boldsymbol{\kappa}^{(R)} = \boldsymbol{\kappa}^{(F)}$ or $\boldsymbol{\Psi}^{(R)} = \boldsymbol{\Psi}^{(F)}$. Indeed, the purpose of an invariance analysis is to determine whether an observed group difference can safely be attributed to a generalizable difference in the distribution of the traits targeted for measurement (common factors) rather than a bias or idiosyncratic property of the particular test.

One can use the likelihood ratio test to assess the statistical significance of a less stringent invariance model's improved fit to the observed means and covariances. However, because psychometric models are highly idealized approximations, a more restrictive model will almost always be rejected if the two samples are sufficiently large. We are thus forced to consider whether the statistically significant discrepancies between groups are quantitatively large enough to merit the exclusion of the offending tests or perhaps even the entire battery. For instance, whereas a difference in standardized factor loadings of 0.05 might be considered negligible, we would certainly be concerned by an intercept difference amounting to several items correct.

A popular approach to such numerical assessment of model adequacy is to summarize the discrepancies with a scalar index and compare its realized value with recommended benchmarks for "poor," "fair," or "good" fit

Table 1 ■ Some Models of Measurement Invariance

Model	Description
configural invariance	$\boldsymbol{\Lambda}^{(R)}$ and $\boldsymbol{\Lambda}^{(F)}$ have the same number of columns and pattern of nonzero elements
weak factorial or *metric* invariance	$\boldsymbol{\Lambda}^{(R)} = \boldsymbol{\Lambda}^{(F)}$
strong factorial or *scalar* invariance	$\boldsymbol{\Lambda}^{(R)} = \boldsymbol{\Lambda}^{(F)}$, $\boldsymbol{\mu}^{(R)} = \boldsymbol{\mu}^{(F)}$
strict factorial invariance	$\boldsymbol{\Lambda}^{(R)} = \boldsymbol{\Lambda}^{(F)}$, $\boldsymbol{\mu}^{(R)} = \boldsymbol{\mu}^{(F)}$, $\boldsymbol{\Theta}^{(R)} = \boldsymbol{\Theta}^{(F)}$

Note. Superscripts are used to denote the group (reference or focal) described by each parameter. Millsap (2011) provides a more complete tabulation of invariance models.

This approach has been criticized for forcing the inherently multifaceted aspects of model-data fit into a single number (McDonald, 2010). Here we accept this criticism and emphasize methods for assessing fit at the level of individual indicators rather than the model as a whole. In our implementation of this approach, prominence is given to numerical indices with straightforward interpretations and graphical methods that compactly convey many pieces of information at once. In this way we aim to dispel the notion that "for purposes of publication, graphical displays are often limited to illustrative examples" (Steinberg & Thissen, 2006, p. 406).

Nonlinear Factor Analysis of Binary Items and Item Response Theory

Factor analysis has long been applied to individual items, which can be regarded as tests yielding just two possible scores: zero (wrong) and one (right). It has also been recognized that the linear factor model, strictly speaking, cannot be appropriate for this purpose (e.g., Carroll, 1945). In the case of a single common factor ($q = 1$), the expected item score becomes inadmissible once the straight regression line either sinks below zero or rises above one for examinees with extreme values of θ. In fact, this kind of problem also arises in the factor analysis of entire tests, each of which yields integer-valued scores bounded between zero and the number of items. The failure of the linear model at the extremes, however, is more pressing when the units of analysis are individual items that are very easy or difficult. In this case the "extremes" are no longer so far from the typical examinee.

A nonlinearization of factor analysis addressing this difficulty must bound the expected item score (probability of giving the correct response) between zero and one, and the simplest such extension seems to be the following (Christoffersson, 1975; Muthén, 1978). Suppose that the jth item is associated with an underlying quantitative response tendency Y_j^* and a threshold τ_j such that the observed item score

$$y_j = \begin{cases} 1 & \text{if } y_j^* > \tau_j, \\ 0 & \text{if } y_j^* \leq \tau_j. \end{cases}$$

Now suppose that $\mathbf{y}^* \in \mathbb{R}^p$, the vector of response tendencies, fits the factor model

$$\mathbf{y}^* = \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\epsilon}^*. \tag{4}$$

Equation 4 lacks the intercept term because we may standardize each Y_j^* so that its mean is zero and its variance unity. If we let $\boldsymbol{\lambda}_j$ stand for the jth row of $\boldsymbol{\Lambda}$ (the jth item's loadings on the q common factors), then we can write the variance of the residual ϵ_j^* as

$$\mathrm{Var}\left(\epsilon_j^*\right) = \mathrm{Var}\left(Y_j^*\right) - \mathrm{Var}\left(\boldsymbol{\lambda}_j'\boldsymbol{\theta}\right) = 1 - \boldsymbol{\lambda}_j'\boldsymbol{\Psi}\boldsymbol{\lambda}_j.$$

Now let us assume that each ϵ_j^* follows the normal distribution in the examinee population. Then

$$\begin{aligned} \mathbb{E}(Y_j \mid \boldsymbol{\theta}) &= \mathbb{P}\left(\boldsymbol{\lambda}_j'\boldsymbol{\theta} + \epsilon_j^* > \tau_j\right) \\ &= \mathbb{P}\left(\epsilon_j^* > \tau_j - \boldsymbol{\lambda}_j'\boldsymbol{\theta}\right) \\ &= \mathbb{P}\left(-\epsilon_j^* \leq \boldsymbol{\lambda}_j'\boldsymbol{\theta} - \tau_j\right) \\ &= \mathbb{P}\left(Z \leq \frac{\boldsymbol{\lambda}_j'\boldsymbol{\theta} - \tau_j}{\sqrt{1 - \boldsymbol{\lambda}_j'\boldsymbol{\Psi}\boldsymbol{\lambda}_j}}\right), \end{aligned}$$

where Z is the standard normal random variable (which has a symmetrical probability distribution). The nonlinear regression of the item score on $\boldsymbol{\theta}$ can thus be written as

$$\mathbb{E}(Y_j \mid \boldsymbol{\theta}) = \Phi\left(\frac{\boldsymbol{\lambda}_j'\boldsymbol{\theta} - \tau_j}{\sqrt{1 - \boldsymbol{\lambda}_j'\boldsymbol{\Psi}\boldsymbol{\lambda}_j}}\right), \tag{5}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. This normal-ogive regression function is monotonically increasing in each element of $\boldsymbol{\theta}$ and bounded between zero and one. The graph

of the item regression function is known as the *item characteristic curve* (ICC), and henceforth we use this term to refer to any item regression function.

For obvious reasons we call Equation 5 the *factor-analytic parameterization* of the ICC. It is sometimes convenient to use the alternative parameterization

$$\begin{aligned} \mathbb{E}(Y_j \,|\, \boldsymbol{\theta}) &= \Phi\left(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{\theta}\right), \\ \tau_j &= \frac{-\alpha_j}{\sqrt{1 + \boldsymbol{\beta}_j' \boldsymbol{\Psi} \boldsymbol{\beta}_j}}, \\ \boldsymbol{\lambda_j} &= \frac{\boldsymbol{\beta}_j}{\sqrt{1 + \boldsymbol{\beta}_j' \boldsymbol{\Psi} \boldsymbol{\beta}_j}}, \end{aligned} \tag{6}$$

which we call the *IRT parameterization* for reasons that we now give. IRT, as formulated by Lord and Novick (1968), is based on the model

$$\begin{aligned} \mathbb{E}(Y_j \,|\, \theta) &= \Phi\left[a_j(\theta - b_j)\right] \\ &\approx \frac{1}{1 + \exp\left[-1.701 a_j(\theta - b_j)\right]}. \end{aligned} \tag{7}$$

The logistic form of Equation 7 proves to be mathematically convenient for computing errors of measurement. In the case of a single factor, if we equate the normal-ogive expression for $\mathbb{E}(Y_j \,|\, \theta)$ in Equation 6 to Lord's normal-ogive form of Equation 7, then we find that β_j is equivalent to a_j (the *discrimination parameter*) and α_j to $-a_j b_j$ (where b_j is called the *difficulty* or *location parameter*). This step shows that there is a sufficiently close correspondence between (α_j, β_j) and Lord's (b_j, a_j) to justify the term "IRT parameterization" for Equation 6. And in fact, it does more; we have now demonstrated the complete equivalence of parametric IRT and the factor analysis of binary measurements.

The $\boldsymbol{\theta}$ in Equation 4 satisfy the principle of local independence—the same principle that serves as a defining property of the $\boldsymbol{\theta}$ in Equation 1. For Y_j^* must be uncorrelated with $Y_{j'}^*$ for each pair j and j', in a subpopulation with no variation in $\boldsymbol{\theta}$, and therefore the fact that an examinee in this subpopulation passed item j ($y_j = 1$ implying $y_j^* > \tau_j$) provides no information as to whether the examine also passed item j' (since the event $y_{j'}^* > \tau_{j'}$ to yield $y_{j'} = 1$ retains its conditional probability given $\boldsymbol{\theta}$ alone). It follows that the adequacy of a given IRT model can be tested by determining whether the off-diagonal elements of the residual covariance matrix tend to vanish (Stout, 1990; McDonald & Mok, 1995). A Bayesian model-checking method along these lines has been given by Levy, Mislevy, and Sinharay (2009). It also follows that the RMSR and GFI carry over as global fit indices from the factor analysis of tests to the IRT analysis of items. Furthermore, Maydeu-Olivares and Joe (2005) have shown that the RMSEA and related indices can be calculated from the fit of the IRT model to the item covariances in a manner analogous to the long-standing practice of factor analysis.[2]

McDonald (1981, 1999, 2013) has forcefully emphasized the close connection between IRT and traditional factor analysis, pointing to the principle of local independence as a unifying theme. However, when other writers have noted this connection, they have done so in passing and without employing it in their further treatment of IRT. This reticence may be owed in part to the different ways in which these two tools have been used and conceptualized throughout their histories.

For the most part, the pioneers of factor analysis regarded the technique as an exploratory tool of natural science and hoped that its judicious application would uncover objective taxonomies of behavioral variation. Ever since the work of Thurstone (1938), many psychologists have embraced the application of multidimensional models ($q \geq 2$) to batteries of instruments that are deliberately heterogeneous in content. Such heterogeneity is perceived as a virtue, since Nature is expected to be complex. Now compare this outlook with the one adopted by the typical IRT practitioner in an applied setting. The object of analysis is usually a single test designed for a narrow educational purpose, and this invites the treatment of tests as parochial artifices rather than measurements of natural kinds. Unidimensionality ($q = 1$) often does hold to a good approximation in IRT applications, which is an unsurprising consequence of testing a narrow range of content. Confinement to one dimension—and the high accuracy with which a long test can place individuals along that dimension—eliminates at one stroke the controversies over underfactoring, overfactoring, rotational indeterminacy, and score indeterminacy that have bestrewn the history of factor analysis without any parallels in the development of IRT.

Some writers have sketched a middle way between these two attitudes (McDonald, 1985, 2003; J. J. Lee, 2012a, 2012b). Briefly, even if factor analysis is limited in its potential for unsupervised discovery, it may still prove to be a useful tool of pure research for those who are willing to accept that *a priori* low-dimensional trait domains have a legitimate place in scientific models. If widely and explicitly adopted, this position may narrow the cultural divide between test-based factor analysis and IRT, thereby removing a hindrance to the use of their formal parallelism.

Purely technical points are sometimes raised against

[2] Because we wish to deprecate exclusive reliance on these scalar fit indices (RMSR, GFI, RMSEA), we do not explain them in detail here. Explications of these indices can be found elsewhere (e.g., Millsap, 2011).

the unification of factor analysis and IRT. For instance, Millsap (2011) points out that a generalization of the IRT model in Equation 7 that allows a nonzero lower asymptote,

$$\mathbb{E}(Y_j \mid \boldsymbol{\theta}) = c_j + (1 - c_j)\,\Phi\left(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{\theta}\right), \tag{8}$$

is no longer in correspondence with a factor-analytic model of the form assumed by Equations 1 and 4. The *pseudo-guessing parameter*, c_j, is often used to model a multiple-choice item to which low-ability examinees have a greater-than-zero probability of giving the correct response. This argument against placing the analysis of tests and items within a common framework seems rather weak because, as pointed out earlier, the linear factor model also requires care when applied in the presence of floor and ceiling effects.

Description of the Dataset

Here we describe the dataset used to illustrate the manner in which a unified factor-analytic framework encompassing IRT can inform the detection of both item- and test-level measurement bias. Note that we are unable to release this dataset; a simulated dataset based on the inferred parameters and `R` code implementing the described analyses can be downloaded from the journal website.

Data from a statewide mathematics achievement test were gathered from roughly 70,000 examinees. Approximately 8 percent were allowed to use their calculators during the administration of the test. There was no limitation on the types of calculators that could be used (e.g., graphing calculators were permitted). We will refer to the accommodated students as the *focal* group. We have not been able to determine whether a standardized diagnostic procedure was used in all cases of claimed disability.

The test consisted of 34 binary items falling into five content categories: Algebra (11 items), Number Sense (10 items), Measurement (3 items), Geometry (3 items), and Statistics (7 items). We have renumbered the items so that those within the same content category are adjacent. The summary statistics are as follows: reference group (no calculator), $M = 22.9$, $SD = 8.0$; focal group (calculator), $M = 13.4$, $SD = 6.0$. It appears that the test as a whole was markedly more difficult for members of the focal group, despite their access to calculators. Broadly speaking, the purpose of our DIF analysis is to determine whether the observed group difference reflects a difference in mastery of the entire content domain (mathematics achievement) rather than an artifact of any biased items that may have been included in this test.

Figure 1 presents an exploratory plot that the analyst should always examine in some detail. Each panel contains the empirical regression of the item score on the total score obtained on all other items. The independent variable thus excludes the focal item. In some cases there are subtle reasons for including the focal item in the total (Lewis, 1993), but for our purposes it is better to exclude the focal item in order to prevent spurious agreement between groups at the test floor and ceiling. The item-test regressions can be an aid in detecting errors in the scoring key or the absence of a single unambiguously correct answer; items afflicted by such problems will tend to exhibit regressions that fail to be monotonically increasing. The regressions also give a global impression of what to expect in the subsequent model-based analysis. In the case of a single factor, the total score approaches a monotonic transformation of θ as the number of items increases, and thus each item-test regression should approximate the corresponding ICC up to a monotonic transformation of the independent variable. The regressions of the reference and focal groups do not appear to differ dramatically, which might lead us to expect that any measurement bias in this test is small.

Testing the Dimensionality of the IRT Model

A prerequisite of DIF analysis is a close fit of the IRT model to the reference group's data, and perhaps the most important aspect of a factor-analytic model at any level is whether its specified dimensionality (number of factors) satisfies the principle of local independence to a good approximation. If two indicators are strong measures of a second factor in addition to the general factor, then an average group difference in the second factor may lead to the misleading appearance of measurement bias afflicting these two indicators in a model specifying one factor only.

We used the `mirt` package for the `R` computing platform (Chalmers, 2012). We first fit a unidimensional IRT model to the item-level data by marginal maximum likelihood. Following the recommendation of Lord (1980), we constrained each c_j to be equal between groups; the item-test regressions in Figure 1 do not reveal any group differences in lower asymptotes that would suggest a severe violation of this constraint. All other parameters were unconstrained, and each group was referred to its own standardized metric. The resulting estimates of the parameters are given in Table 2.

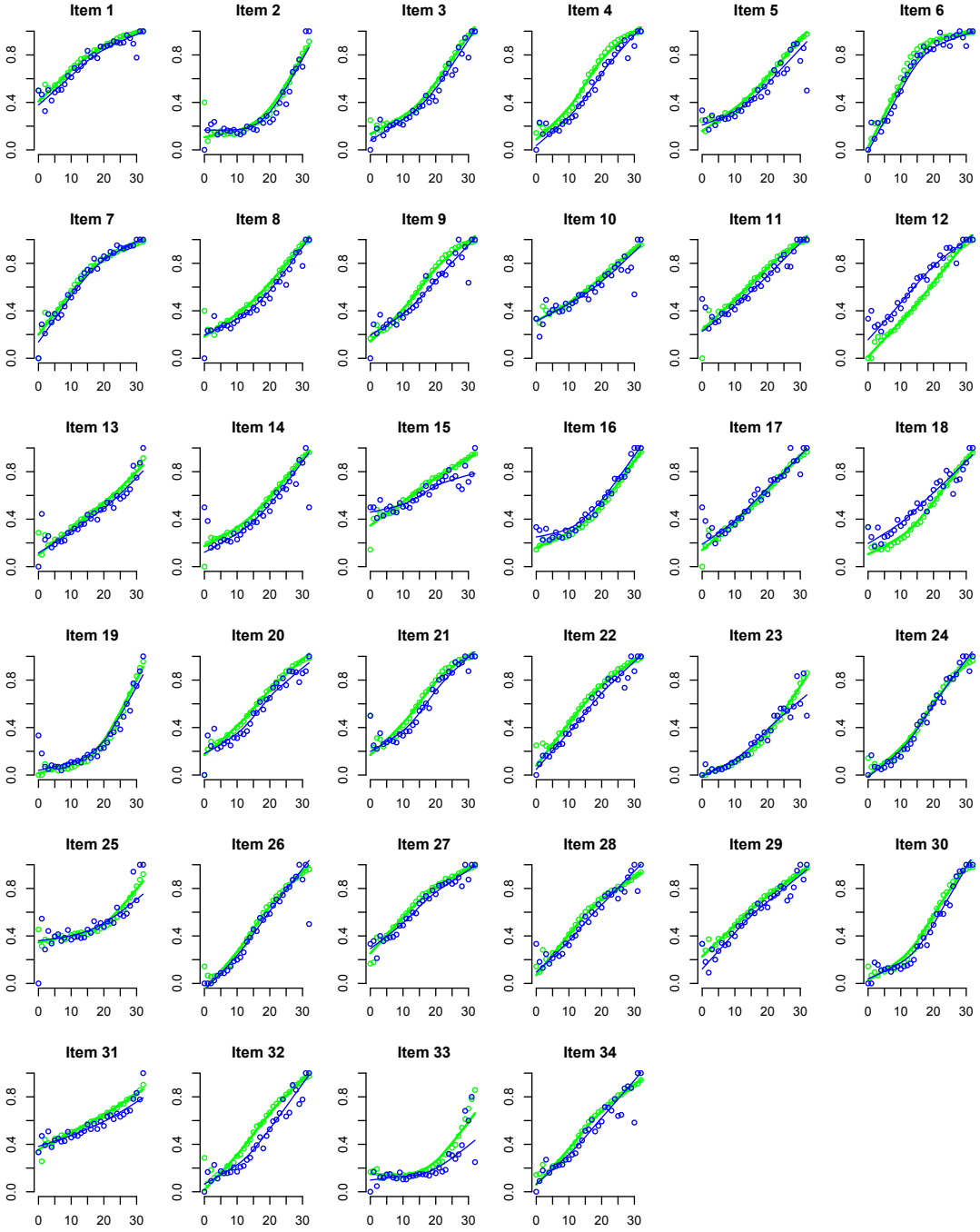
The global fit indices suggest that the unidimensional model fits the data well, $\chi^2(1020) = 18{,}539.54$, RMSEA = 0.022, $\text{RMSR}^{(R)} = 0.0038$, $\text{RMSR}^{(F)} = 0.0051$, $\text{GFI}^{(R)} = 0.9961$, $\text{GFI}^{(F)} = 0.9889$. The RMSEA in the focal group alone was 0.021. But global fit indices can only be a supplement at best to the necessary granular view of model-data fit provided by Figure 2, which gives the matrix of residual covariances (differences between sample and model-predicted covariances). Here the residual covariances have been divided by the sample covariances themselves to aid interpretation.

Table 2 ■ Item Parameter Estimates: Nonlinear Factor-Analytic Parameterization

Item	c	$\tau^{(R)}$	$\tau^{(F)}$	$\lambda^{(R)}$	$\lambda^{(F)}$	$\tau^{(F^*)}$	$\lambda^{(F^*)}$
Algebra							
1	0.434	−0.740	0.233	0.718	0.651	−0.780	0.744
2	0.124	0.302	1.454	0.768	0.699	0.364	0.800
3	0.218	−0.090	1.132	0.840	0.754	−0.042	0.862
4	0.116	−0.590	0.629	0.818	0.705	−0.469	0.806
5	0.229	−0.162	0.911	0.733	0.642	−0.089	0.734
6	0.002	−1.122	−0.157	0.732	0.697	−1.243	0.797
7	0.024	−0.906	−0.178	0.611	0.565	−1.059	0.646
8	0.291	−0.138	1.124	0.801	0.739	−0.027	0.845
9	0.184	−0.548	0.497	0.770	0.584	−0.414	0.669
10	0.335	−0.183	0.697	0.661	0.502	−0.086	0.575
11	0.338	−0.431	0.722	0.762	0.677	−0.333	0.774
Number Sense							
12	0.196	−0.263	0.241	0.808	0.629	−0.738	0.719
13	0.122	−0.056	0.691	0.543	0.464	−0.032	0.531
14	0.197	−0.129	0.967	0.711	0.666	−0.071	0.762
15	0.296	−0.404	0.337	0.555	0.352	−0.212	0.403
16	0.217	0.049	0.897	0.761	0.662	−0.136	0.758
17	0.202	−0.261	0.522	0.684	0.602	−0.416	0.689
18	0.114	−0.129	0.399	0.717	0.472	−0.335	0.539
19	0.043	0.274	1.277	0.801	0.626	0.303	0.716
20	0.211	−0.431	0.673	0.743	0.690	−0.402	0.789
21	0.241	−0.571	0.660	0.836	0.774	−0.546	0.885
Measurement							
22	0.104	−0.627	0.358	0.689	0.601	−0.579	0.687
23	0.002	0.153	0.906	0.643	0.588	−0.011	0.673
24	0.009	−0.348	0.592	0.732	0.675	−0.461	0.773
Geometry							
25	0.380	0.477	1.457	0.760	0.675	0.406	0.771
26	0.002	−0.426	0.502	0.693	0.653	−0.515	0.747
27	0.185	−0.723	0.125	0.628	0.532	−0.703	0.608
Statistics							
28	0.004	−0.544	0.218	0.539	0.501	−0.563	0.573
29	0.102	−0.584	0.205	0.564	0.482	−0.546	0.552
30	0.078	−0.253	0.966	0.827	0.759	−0.217	0.868
31	0.316	−0.070	0.611	0.469	0.325	0.105	0.371
32	0.043	−0.439	0.654	0.695	0.554	−0.210	0.634
33	0.119	0.576	1.915	0.768	0.742	0.758	0.849
34	0.005	−0.481	0.322	0.578	0.475	−0.418	0.543

Note. Superscripts are used to denote the group (reference or focal) described by each parameter.

Figure 1 ■ Each panel displays the proportion of examinees giving the correct response as a function of the total score on all other items. The LOESS curve is also plotted. Green (light shading) corresponds to the reference group (no calculator), whereas blue (dark shading) corresponds to the focal group (calculator). The content strands are Algebra (items 1 to 11), Number Sense (12 to 21), Measurement (22 to 24), Geometry (25 to 27), and Statistics (28 to 34).



Despite the recognized usefulness of residual covariances in assessing factor-analytic models applied to test data, they are less often used in assessing IRT models applied to item data—perhaps because the covariance between binary items predicted by their model parameters is no more readily computed from those parameters than other measures of residual dependence that have been explored. J. J. Lee and Lee (2016) review a method for approximating the normal-ogive ICC that allows the expected covariance between items to be readily computed. Although this computation is based on the principles underlying the NOHARM program (Fraser & McDonald, 1988), it can be applied to items whose parameters have been estimated by any method. We used this approach here to calculate the residual covariance between each pair of items.

The lower off-diagonal elements of the residual covariance matrix constitute a nearly uniform sea of green (intermediate shading), which shows that the model fits the reference group's data extraordinarily well. Over 50 percent of the reference group's normalized residual covariances are between −0.053 and 0.032, and about 90 percent are between −0.133 and 0.130. The smallest and largest normalized residuals are −0.197 and 0.281 respectively; in other words, not one of the 561 positive covariances is missed by as much as 30 percent, and most of the misses are far smaller.

The vast majority of the focal group's residuals are also negligible (green or intermediate shading), but there are noticeably more residuals that are large as a percentage of their target covariances. There are several reasons, however, not to take these relatively large residuals as evidence for a substantial violation of configural invariance (Table 1). The focal group is plausibly characterized by smaller variance in θ, and smaller variance in the dominant factor has the effect of magnifying any "nuisance" factors whose variances are not similarly diminished. How well the approximation of unidimensionality holds is thus somewhat dependent on the population even in the absence of measurement bias (Reckase, 2009). When the goal is to refer all examinees to the metric of a highly variable population, the greater prominence of nuisance factors in a less variable subgroup is not necessarily a cause for concern.

The term "nuisance" implies a lack of systematicity—that the content of the items departing from local independence provides no guidance as to how we might write additional items that deliberately measure or avoid the additional factors. This criterion is obviously violated if items tend to exhibit large residual covariances with other items in their content category. Because we have numbered the items so that those within the same content category are contiguous, substantively meaningful violations of unidimensionality attributable to content categories should lead to a clustering of the larger positive residuals near the diagonal of the focal group's matrix in Figure 2. Any such clustering, however, is not visually obvious.

As a further check, we fit an exploratory two-factor model (Promax rotation) by the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010). Even after fixing the pseudo-guessing parameters to zero in order to alleviate a convergence problem, the fit indices specific to the focal group did improve, $\text{RMSR}^{(F)} = 0.004$, $\text{GFI}^{(F)} = 0.9932$. There was no discernible pattern in the factor loadings, however, as items within the same content category did not tend to exhibit stronger loadings on one factor rather than the other. The estimated correlation between the factors in the focal group was 0.72, which should probably be regarded as a lower bound; typically, fixing the non-salient loadings to zero in a confirmatory model leads to an increase in the estimated factor correlation with negligible loss of fit. The non-generalizable nature of these two factors and their high correlation both weigh strongly in favor of accepting a unidimensional model and proceeding to an examination of measurement bias.

Rescaling of Focal Parameters to the Reference Metric

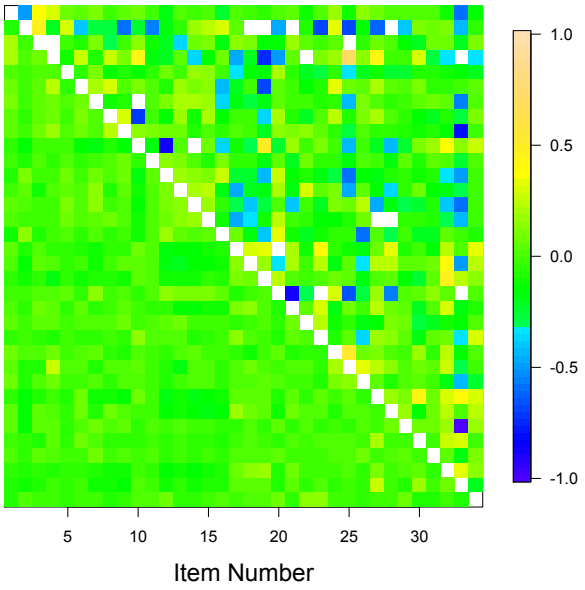
At this point many factor analysts might expect a procedure that has become somewhat *de rigeur* in the psychological literature: the imposition of each successive constraint in the middle two rows of Table 1, substituting the thresholds $\boldsymbol{\tau}$ for the intercepts $\boldsymbol{\mu}$, and the determination of whether model fit deteriorates substantially. (The residual variance is no longer independent of the conditional expectation in the case of a binary variable. Strict factorial invariance is thus no longer distinct from strong invariance and need not be tested.) Since even trivial group differences in model parameters will be significant in large enough samples, changes in fit indices such as the RMSEA are used in practice to aid the judgment of whether substantial bias is present.

This procedure is perhaps useful in some cases, and its application to our dataset will later be illustrated. But first we will demonstrate a different approach, set out in part by Lord (1980) and McDonald (1999), that is more in line with this article's emphasis on graphics and granularity.

The parameter estimates will differ between groups as a result of their different distributions of θ, since in each group θ is standardized with respect to its own mean and standard deviation. We require a transformation of the focal group's parameters that—in the absence of measurement bias—will bring them into agreement with those of the reference group. The scales of the two groups are related by the transformation

$$\theta^{(F)} = u + v\theta^{(R)}, \tag{9}$$

Figure 2 ■ The residual covariance matrix, where each element is proportional to the difference between the sample covariance and the covariance predicted by the best-fitting IRT parameters. The difference is divided by the sample covariance itself to aid interpretation. The reference group (no calculator) is represented by the lower triangle, while the focal group (calculator) is represented by the upper triangle. The progression of the color scale from blue (dark shading) to peach (light shading) corresponds to the progression from −1 to +1. The diagonal has been set to white for clarity. Any absolute normalized residual covariance exceeding 100 percent has also been set to white.



which has the inverse

$$\theta^{(R)} = \frac{1}{v}\theta^{(F)} - \frac{u}{v}. \tag{10}$$

The denominator of the argument in Equation 5 is the residual standard deviation, the square root of $\mathrm{Var}\left(\epsilon_j^*\right)$, whose equality between groups is assured by the equivalent of scalar invariance. Group invariance of the function in Equation 5 thus implies the equality of the numerator. Express the numerator in focal units,

$$\lambda_j^{(F)}\theta_j^{(F)} - \tau_j^{(F)},$$

and then note that the conversion of θ to reference units using Equation 10 requires the substitutions of

$$\begin{aligned} \lambda_j^{(F^*)} &= v\lambda_j^{(F)} \\ \tau_j^{(F^*)} &= \tau_j^{(F)} - \lambda_j^{(F)}u, \end{aligned} \tag{11}$$

for $\lambda_j^{(F)}$ and $\tau_j^{(F)}$ if the numerator is to retain its numerical value. Equation 11 thus gives the required transformations of the focal parameters, which are equal to $\lambda_j^{(R)}$ and $\tau_j^{(R)}$ in the absence of DIF.

We seek the values of u and v that bring the reference and rescaled focal parameters as close together as possible. It is straightforward to show that

$$\sum_j \left(\tau_j^{(R)} - \tau_j^{(F^*)}\right)^2 = \sum_j \left[\tau_j^{(R)} - \left(\tau_j^{(F)} - \lambda_j^{(F)}u\right)\right]^2$$

attains its minimum when

$$u = \frac{\sum_j \left(\tau_j^{(F)} - \tau_j^{(R)}\right)\lambda_j^{(F)}}{\sum_j \left(\lambda_j^{(F)}\right)^2}. \tag{12}$$

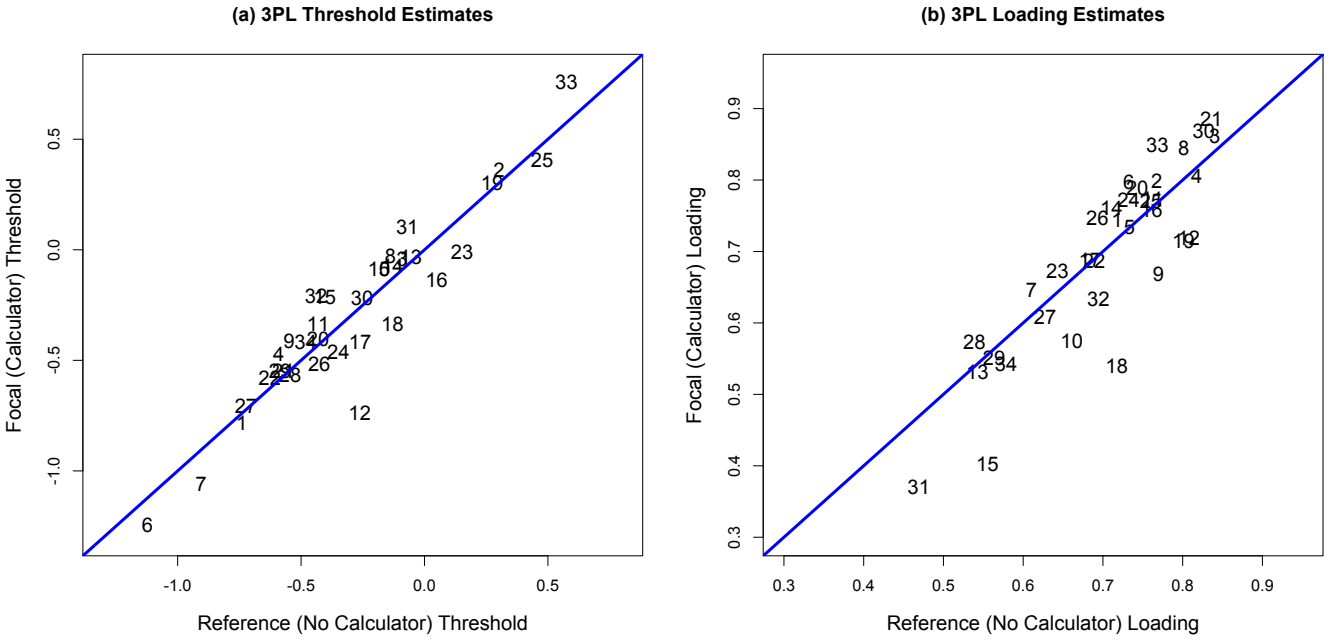
Similarly,

$$\sum_j \left(\lambda_j^{(R)} - \lambda_j^{(F^*)}\right)^2 = \sum_j \left(\lambda_j^{(R)} - v\lambda_j^{(F)}\right)^2$$

attains its minimum when

$$v = \frac{\sum_j \lambda_j^{(R)}\lambda_j^{(F)}}{\sum_j \left(\lambda_j^{(F)}\right)^2}. \tag{13}$$

Figure 3 ■ In each panel the line of zero intercept and unit slope is superimposed. The parameters of the focal group (calculator) have been rescaled to the origin and metric of the reference group (no calculator). The coordinates of each point can be found in Table 2. (a) The scatterplot of the thresholds (τ) in the two groups when the pseudo-guessing parameters are constrained to be equal between groups but otherwise are freely estimated. (b) The corresponding scatterplot of the factor loadings (λ).



These estimates of u and v can be obtained by summing only over items believed to be free from DIF, although in practice this precaution will often scarcely affect the results. For simplicity, the entries in the rightmost two columns of Table 2 were calculated using all items to estimate u and v.[3]

Figure 3 plots the estimates given in Table 2. In each panel the x-axis corresponds to the parameters of the reference group and the y-axis to the rescaled parameters o the focal group. In the absence of DIF, the points in each panel should lie close to the line of zero intercept and unit slope; the graph of this line is superimposed on each panel. Items 12 through 21 are in the Number Sense category, and several of them are among the items with threshold parameters deviating downward from the line in Figure 3a. If we were to discount these items, then the line would indeed pass very close to most of the remaining data points. Turning to the factor loadings in Figure 3b, we see several items (many from the Number Sense category again) lying well below the line. Discounting these items would also bring the line close to the majority of the data points.

It may be possible to refine the rescaling of the focal group's items parameters by using the entries in the rightmost two columns of Table 2 as starting values for an iterative minimization procedure (Stocking & Lord, 1983). We can see from the nearness of the data points to the straight lines in Figure 3, however, that the rescaling already appears to be quite successful. It has been known for some time that applying close analogues of the rescaling procedure employed here to the IRT parameterization often fails to yield satisfactory results (e.g., Kolen & Brennan, 2014). The factor-analytic parameterization appears to support numerically stable rescaling because a factor loading is bound between zero and one (whereas Lord's a

[3] For brevity, we omit all standard errors from our tabulated results. They are available upon request. We calculated the standard errors of the parameter estimates in Table 2 using two methods: (1) inverting the information matrix, which was based on the variance of the Fisher scores, and (2) using 500 replicates of the delete-($20\sqrt{n}$) jackknife (Efron & Tibshirani, 1993). The first method returned much larger estimates of the errors in the estimation of the pseudo-guessing parameters. In the case of such a discrepancy, we recommend non-parametric resampling methods such as the jackknife. We used our jackknife-based standard errors to calculate the significance of the differences in Table 3.

Table 3 ■ Quantification of Differential Item Functioning

Item	$\tau^{(R)} - \tau^{(F^*)}$	$\lambda^{(R)} - \lambda^{(F^*)}$	I_1	I_2
Algebra				
1	0.040	−0.026	0.000	0.007
2	−0.062*	−0.032	0.014	0.014
3	−0.048	−0.022	0.014	0.014
4	−0.121*	0.012	0.033	0.033
5	−0.073*	−0.001	0.016	0.016
6	0.120*	−0.064*	−0.016	0.034
7	0.153*	−0.036	−0.044	0.044
8	−0.111*	−0.043	0.030	0.030
9	−0.134*	0.101	−0.007	0.032
10	−0.098*	0.086	−0.008	0.017
11	−0.097*	−0.013	0.027	0.027
Number Sense				
12	0.476*	0.089*	−0.183	0.183
13	−0.024	0.012	0.002	0.003
14	−0.058*	−0.051	0.031	0.031
15	−0.192*	0.152*	−0.007	0.032
16	0.185*	0.003	−0.036	0.036
17	0.156*	−0.005	−0.042	0.042
18	0.206*	0.178*	−0.146	0.146
19	−0.029	0.085	−0.018	0.021
20	−0.029	−0.046	0.029	0.029
21	−0.025	−0.049	0.029	0.032
Measurement				
22	−0.048	0.002	0.017	0.017
23	0.163*	−0.030	−0.028	0.028
24	0.112*	−0.040	−0.017	0.021
Geometry				
25	0.071*	−0.011	−0.003	0.003
26	0.089*	−0.054	−0.002	0.021
27	−0.020	0.019	−0.002	0.006
Statistics				
28	0.020	−0.034	0.012	0.015
29	−0.038	0.012	0.008	0.008
30	−0.035	−0.042	0.026	0.027
31	−0.175*	0.098	0.007	0.017
32	−0.230*	0.061	0.047	0.047
33	−0.183*	−0.080	0.018	0.018
34	−0.063*	0.036	0.005	0.012
Total			−0.197	1.066

Note. An asterisk indicates $p < .001$. We calculated p-values on the assumptions of exact rescaling and normal sampling distributions.

Table 4 ■ Fit Measures of Item-Based Invariance Models

Model	df	M_2	Δdf	ΔM_2	p	RMSEA	AIC
configural	1,020	18,539				0.0216	2,552,336
metric	1,052	18,582	32	42	> .100	0.0213	2,552,525
scalar	1,086	18,796	34	214	< .001	0.0211	2,554,067
partial (free 12, 15, 18)	1,080	18,599	−6	−197	< .001	0.0210	2,552,840
partial (free 12, 15, 18, 31–34)	1,072	18,559	−8	−40	< .001	0.0211	2,552,745

Note. The Akaike information criterion (AIC) is another commonly used index that, like the RMSEA, attempts to balance model-data fit and parsimony. Smaller values are supposed to correspond to better fit. Whereas the other fit indices are derived from the M_2 statistic of Maydeu-Olivares and Joe (2005), the values of the AIC are derived from the full-information likelihood.

parameter is potentially unbounded) and a threshold is a simple transformation of the pass rate.

Rescaling of the Focal Distribution to the Reference Metric

Figure 3 highlights a number of items departing far from the straight lines. But can we conclude that these items are functioning differentially? Theorists have pointed out that an unbiased item can nevertheless present a spurious appearance of DIF if multiple triplets of parameter values are able to model an ICC well over a small part of its range. For this reason some of these theorists recommend turning to non-IRT tools for bias detection such as logistic regression (e.g., Hambleton, Swaminathan, & Rogers, 1991). In our view, however, the rapidity of plotting and numerical integration enabled by modern computing power allows practitioners to address the possible unstable estimation of group-invariant IRT parameters and thus continue to study bias within a unified factor-analytic framework. We will now substantiate this claim.

If a particular item's bias is spurious in the sense just described, then the graphs of the reference and focal ICCs will be globally discrepant but coincide reasonably well over the interval of θ where the bulk of the focal examinees reside. Drawing these graphs requires expressing the focal distribution of θ in terms of the reference metric. According to Equation 10, the focal value of zero is mapped to $-(u/v)$, and this is therefore the focal mean when the origin and unit are given by the reference mean and standard deviation respectively.

If we were working with a linear factor model, we would apply Equation 12 and 13 to the intercepts and loadings in Equation 1. We could then compute the focal variance in terms of the reference metric as follows. Consider that a focal factor loading in a linear model gives the ratio

$$\frac{\text{number correct}}{\text{focal SD}(\theta)}$$

and that multiplication of this by v must give the ratio

$$\frac{\text{number correct}}{\text{reference SD}(\theta)}.$$

It follows that v must be the ratio of the reference and focal standard deviations.

The straightforward estimate $1/v$ of the focal standard deviation does not apply in a parametric IRT model because a focal factor loading now gives the ratio

$$\frac{\text{focal SD}\,(Y^*)}{\text{focal SD}\,(\theta)},$$

the numerator of which does not appear in the rescaled loading, and thus v can no longer be the ratio of the reference and focal standard deviations with respect to the common factor θ.

J. J. Lee and Lee (2016) show that a least-squares linear approximation of item j's ICC is given by

$$\begin{aligned} \mathbb{E}(Y_j \mid \theta) &\approx \gamma_j + \delta_j \theta, \\ \gamma_j &= c_j + (1 - c_j)\Phi(-\tau_j), \\ \delta_j &= (1 - c_j)\lambda_j \phi(\tau_j), \end{aligned} \tag{14}$$

where Φ and ϕ respectively denote the cumulative probability and density functions of the standard normal distribution. Equation 14 is thus the linear factor model that best approximates the nonlinear model represented by Equation 8. If each ICC closely followed this approximation, we could estimate the focal mean and variance using the values of u and v obtained from the application of Equations 12 and 13 to the intercepts (γ_j) and loadings (δ_j).

Applying this procedure to our dataset, we estimated the focal group's mean to be −1.36 and its standard deviation to be 0.80. We used these estimates to plot the reference and focal ICCs in Figure 4. More specifically, we plotted the focal ICC over the interval $-1.36 \pm 2(0.80)$ and compared this segment to the reference ICC over the same interval. We are particularly interested in whether the items that appear to be afflicted by DIF in Figure 3 exhibit truly

discrepant ICCs over this interval. Failure of the ICCs to agree outside of this interval may simply reflect the inability of the focal group's responses to reflect the nonlinear behavior of the ICC far from where the focal distribution allocates the most probability.

We admit that this graphical method does not yet possess a convenient multidimensional generalization. An important numerical complement to visual comparison is thus to compute the total impact on the average score of focal examinees. For this purpose Wainer (1993) introduced two indices of impact,

$$I_1 = \int \left[\mathbb{E}^{(R)}(Y_j \,|\, \boldsymbol{\theta}) - \mathbb{E}^{(F)}(Y_j \,|\, \boldsymbol{\theta})\right] g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{15}$$

and

$$I_2 = \int \left|\mathbb{E}^{(R)}(Y_j \,|\, \boldsymbol{\theta}) - \mathbb{E}^{(F)}(Y_j \,|\, \boldsymbol{\theta})\right| g(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{16}$$

where $g(\boldsymbol{\theta})$ is an estimate of the focal density function. I_1 is the reduction in the mean score obtained by the focal group as a result of any bias in the item; a negative value indicates that the focal group actually benefits from the differential functioning. A problem with this numerical measure of impact is that it can assume a small value even in the case of substantial *nonuniform bias*: an intersection of ICCs that favors the focal group on one side of the crossover point and the reference group on the other. To avoid this cancellation of opposing biases, the I_2 measure integrates the absolute value of the difference between reference and focal ICCs over the focal distribution of θ.

To calculate I_1 and I_2 in our own application, we assumed that the focal distribution is normal. Table 3 gives the group differences in factor-analytic item parameters and also the two measures of impact.

With all of this machinery in place and its products displayed in tabular and graphical form, we are now ready to single out items for excessive DIF. It is perhaps worth noting at the outset that the sum of I_1 over all items—a measure of the test-wide signed bias favoring the reference group—is actually *negative* and roughly equal in magnitude to a fifth of a point (Table 3). Differential functioning within this test does not seem to hinder members of the focal group and may actually favor them very slightly.

Items in the Number Sense Content Category

Figure 3a suggests that item 12 exhibits threshold bias *favoring* the focal group, and this impression is affirmed by the graphs of the ICCs in Figure 4. The value of I_1 indicates that the bias has brought the mean test score of the focal group closer to that of the reference group by nearly a fifth of a point.

Figure 3b suggests that item 18 exhibits slope bias such that it is a worse indicator of the common factor in the focal group, and the loading difference of 0.178 is large by traditional factor-analytic standards. The graphs of the ICCs in Figure 4 show that this bias is a nearly uniform one *favoring* the focal group; since the region where the item discriminates best is to the right of most focal examinees, a flatter slope effectively raises the focal ICC above the reference ICC. The value of I_1 indicates that the DIF of item 18 has brought the mean test score of the focal group closer to that of the reference group by about a seventh of a point.

Figure 3b suggests that the next most problematic instance of slope bias is exhibited by item 15. The graphs of the ICCs in Figure 4 show that this bias is nonuniform; this conclusion can also be drawn from the fact that I_2 is much greater than I_1. To the eye, however, the total impact of this bias seems rather small. We can also see in Table 3 that items 12 and 18 exert far more of an impact (indexed by either I_1 or I_2) than any other in the entire test.

Item 12 asks the examinee for the integers that bound $\sqrt{14}$ on the number line. Item 18 asks the examinee for the value of the expression $11 \times 2 + \sqrt{25}$. It seems quite plausible that a calculator will make these items easier. In ability and achievement testing, the causes of DIF are often difficult to discern (O'Neill & McPeek, 1993), and it is thus gratifying that in our application we can find a reasonably clear connection between the content of the most differentially functioning items and the nature of the distinction between the examinee populations. Whether the bias favoring the focal group amounts to an unfair advantage depends on the motivation of the state in providing calculators to these examinees and the purposes to which it puts the test scores. Sinharay and Haberman (2014) provide a sophisticated analysis of how the latter consideration can affect judgments regarding the usability of the suspect items or the test as a whole.

Items in the Statistics Content Category

Figure 3b suggests that item 31 might suffer from substantial slope bias reducing its discriminatory power in the focal group. The graphs of the ICCs in Figure 4 show that this bias is a nonuniform one. The impact of this item, as indexed by I_2, is smaller because its factor loading is already rather low in the reference group.

So far we have looked for graphical evidence of DIF in Figure 3 and used Figure 4 to check that such evidence is not a spurious result of the difficulty in estimating a nonlinear ICC from a limited range of θ. Now we will study the numerical evidence in Table 3 and check that our graphical approach has not missed any important patterns. It appears that item 31 is one of several items in the Statistics content category that exhibits threshold bias *disfavoring* the focal group. The final few panels of Figure 4 suggest that the bias may be practically negligible, but its con-

centration within this content category may be of substantive interest. The Statistics items requires the understanding of concepts such as probability, mean, and median, and there is no obvious reason why the provision of calculators would have hindered the focal examinees here. Psychologists studying dyscalculia may consider whether a disproportionate difficulty with such concepts accords with the current understanding of this disability or suggests new avenues of research.

Global Testing of Nested Models

For completeness we now illustrate the fitting of the nested models in Table 1—a common practice in the examination of test-level bias—to the item-level data that we have been analyzing so far. Table 4 shows that the RMSEA and AIC do not agree on the tenability of metric invariance. However, since forcing just one set of parameters to be equal between groups may merely relocate genuine DIF to the other set, it may be sensible to ignore this ambiguity and proceed to a test of scalar invariance (no DIF).

A limited-information test of significance yields a minuscule p-value against the null hypothesis of scalar invariance. The increase in the AIC also points toward a rejection of scalar invariance. But the RMSEA continues to decline and thus suggests that scalar invariance should be accepted. There is perhaps a sense in which this latter conclusion has some merit. Considered as a whole, the test *does* display relatively little measurement bias. But remaining content with this global characterization would lead one to miss the rich implications of our more thoroughgoing analysis. The sole reliance on sequential molar tests of model-data fit therefore cannot be recommended; this procedure should be regarded as complementary to our item-level approach.

Partial invariance refers to a model where a minority of the indicators are allowed to function differentially. To yield such a model, we freed the thresholds and factor loadings of the three Number Sense items singled out in the previous section. This led to further decreases in both the RMSEA and AIC. We specified an even less restrictive model of partial invariance by freeing the Statistics items singled out in the previous section, but this time the RMSEA and AIC were discordant. Some ambiguity is perhaps to be expected in light of the fact that the differential functioning of the last four items is visually discernible in Figure 4 but still quite small.

Of course, the choice of items to be freed was made on the basis of our earlier analysis. How might the choice be made otherwise? A given parameter's *modification index* (MI) is the extent to which the formal test statistic decreases if its value is permitted to vary between groups, and factor analysts often resort to repeatedly freeing the parameter with the largest MI until they reach a model which which they are satisfied. However, without the guidance of the fine-grained results displayed and tabulated in our earlier analysis, there do not seem to be any compelling criteria for whether this greedy procedure has reached a sensible endpoint.

Incidentally, our final model of partial invariance led to an estimate of the focal mean equaling −1.32 and an estimate of the focal standard deviation equaling 0.83. These estimates are in excellent agreement with those obtained by our rescaling procedure.

Implications for Bias Detection at the Test Level

At this point the reader might be wondering whether our broadly negative evaluation of single-valued fit indices is even more applicable to the studies of test-level bias where they are commonly used. After all, the smaller number of indicators should make our alternative approach easier to implement at the test level. This is in fact precisely the position that we adopt: unless accompanied by the indicator-based approach emphasized in this article, the global testing of nested models is an inadequate procedure *regardless* of whether the indicators are single items or entire tests.

Since the global approach is well entrenched in factor-analytic investigations of test-level bias, a demonstration of our indicator-based approach to a battery of tests may be required to convince users of its viability in this context. For this purpose we will treat each content category as a distinct test and step through a multiple-group factor analysis of the test battery formed (artificially) in this way. Measurement bias at the test level can be called *differential test functioning*. Although this term is seldom used in the literature, it is useful because it highlights the close parallel between analyses at the item and test levels.

As expected from our item-level dimensionality analysis, the fit of a single-factor model is outstanding in both the reference and focal groups, $\chi^2(10) = 261.05$, $\text{RMSEA}^{(R)} = 0.027$, $\text{RMSEA}^{(F)} = 0.019$, $\text{RMSR}^{(R)} = 0.007$, $\text{RMSR}^{(F)} = 0.009$, $\text{GFI}^{(R)} = 0.9999$, $\text{GFI}^{(F)} = 0.9998$. Note that we used the correlation matrices of the tests to calculate the RMSR and GFI. No element of either group's residual correlation matrix exceeds 0.02 in absolute magnitude.

Equations 12 and 13 can be applied to express the focal parameters in terms of the reference origin and unit in exactly the same manner as in the item-level analysis, and the results of this rescaling are given in the fifth and sixth columns of Table 5. The reference and rescaled focal parameters are also plotted in Figure 5, the test-level equivalent of Figure 3. As in the item-level analysis, indicators suffering from relatively large measurement bias will lie far from the superimposed straight lines. Figure 5a shows the pattern of intercept bias expected from the item-

Table 5 ■ Test Parameter Estimates: Linear Factor-Analytic Parameterization

Test	$\mu^{(R)}$	$\mu^{(F)}$	$\lambda^{(R)}$	$\lambda^{(F)}$	$\mu^{(F^*)}$	$\lambda^{(F^*)}$	$\sum_j \gamma_j$	$\sum_j \delta_j$
Algebra	8.028	4.836	2.244	1.710	7.909	2.373	8.017	2.215
Number Sense	6.450	3.947	2.302	1.546	6.725	2.145	6.478	2.232
Measurement	1.846	0.893	0.743	0.594	1.961	0.824	1.842	0.728
Geometry	2.055	1.287	0.583	0.474	2.134	0.658	2.051	0.577
Statistics	4.517	2.436	1.472	1.017	4.263	1.411	4.503	1.451

Note. Superscripts are used to denote the group (reference or focal) described by each parameter. For brevity the focal group's $\sum_j \gamma_j$ and $\sum_j \delta_j$ are not shown.

level analysis: Number Sense is characterized by bias favoring the focal group and Statistics by bias favoring the reference group. The quantitative magnitudes of these biases, however, seem to be quite small.

The quantities u and v can be used immediately to calculate the focal mean and standard deviation; we estimated these to be -1.29 and 0.72 respectively. The agreement with the estimates derived from the item-level analysis is reasonably good.

Each entry in the second rightmost column of Table 5 is the sum, over all items in the given test, of the constant terms γ_j in the linear approximations of the reference ICCs (Equation 14). Similarly, each entry in the rightmost column is the sum of the coefficients δ_j in the approximations. Note the very close agreement between these sums, which are derived from an IRT analysis, and the standard test-level factor-analytic intercepts and loadings. The sum of the ICCs over the items in a particular test is called the *test characteristic curve* (TCC) in the IRT literature; it provides the expected number of items answered correctly as a function of the common factor (θ). The results in Table 5 thus remind us of a point rarely made in the factor-analytic literature: a test's intercept and factor loading parameterize an approximation of its TCC.

Figure 6, the test-level equivalent of Figure 4, more clearly depicts the close relationship between the IRT-based and factor-analytic forms of the TCC. The solid lines represent the latter, and they confirm our earlier observations regarding the absence of serious measurement bias. For the reasons given in our item-based analysis, each focal line is plotted only over $-1.29 \pm 2(0.72)$. Number Sense shows the largest discrepancy between the reference and focal TCCs, and the discrepancy favors the focal group. Also plotted in each panel are the TCCs formed from summing, not the linearized ICCs given by Equation 14, but the standard normal-ogive ICCs prescribed by parametric IRT and studied at length in our earlier item-based analysis. These nonlinear TCCs also agree very closely with their linear factor-analytic counterparts over the interval of θ where the majority of the focal examinees are located.

Notice that the linear TCC fails at high values of θ because of a ceiling effect: each test is predicted to yield impossibly high scores in the high-θ region. In principle, such a failure can lead to the spurious appearance of measurement bias, although in our case the failure is inconsequential because most focal examinees are located in the low-θ region where the linear model approximates the TCC well. But an application where there is a floor rather than ceiling effect and the focal group is again less able on average might well produce the misleading appearance of differential test functioning. If the analyst lacks the item-level data needed to provide a direct check of this possibility, another option is to conduct a nonlinear factor analysis of the test-level data (McDonald, 1967). In fact, the normal-ogive form with a lower asymptote seems to model each TCC well over the entire depicted range of θ, and one could adapt the least-squares procedure described in J. J. Lee and Lee (2016) to fit covariances between whole tests after each one's scores are multiplied by the reciprocal of the test length in items.

Once we realize that item and test characteristic curves are the same type of object, it is easy to appreciate test-level equivalents of I_1 and I_2 as useful indicator-based measures of impact or effect size. We used Equations 15 and 16 to calculate the difference between the reference and focal factor-analytic TCCs and obtained results in harmony with our item-level analysis: Number Sense shows a signed bias favoring the focal group by over 0.40 items correct, whereas Statistics shows a signed bias favoring the reference group by nearly 0.20 items correct.

Summary and Conclusion

In this article we set out an account that unifies linear factor analysis (typically applied to a battery composed of several tests) and item response theory (typically applied to a test composed of several items), placing the detection of measurement bias at both levels in a common framework. We then demonstrated several ways in which this important

application of test theory benefits from theoretical parsimony.

Long-established factor-analytic criteria—based on whether residual covariances tend to vanish as prescribed by the principle of local independence—can readily be applied to judge how well a model of fixed dimensionality fits item-level data. It is becoming more common for papers containing linear factor analyses or structural equation models to print the entire residual covariance (correlation) matrix, and our graphical displays show that this practice can be extended to the large matrices typical of item-level data. It also follows that factor-analytic scalar fit indices can be used to asses model-data fit in the IRT setting, although we do not endorse relying solely on such indices in either type of analysis. Once the model has been fit and accepted, the factor-analytic parameterization of IRT is valuable because its properties facilitate interpretative standards inherited from test-level analysis and numerically stable rescaling of the focal group's parameters.

The study of differential *test* functioning may actually stand to gain more from the unified framework than that of differential *item* functioning. While our graphical methods permit an even more comprehensive indicator-based approach, a focus on single items has consistently been emphasized in the IRT literature. By analyzing the same item data (except coarsened to test-level resolution) in exactly the same fashion and reaching essentially the same conclusions, we have highlighted the fact that our approach is general enough to cover test-based analyses. Our recommendation to psychologists studying measurement bias with factor analysis is therefore to supplement the standard global testing of nested invariance models with our graphical and indicator-based methods.

Ignoring the item-level nature of psychological measurements is obviously somewhat artificial and in this article was done mostly for pedagogy. Does this mean that the linear factor analysis of whole tests is an outmoded approach to the detection of measurement bias? Not necessarily. First, in many datasets the item-level scores are unrecorded, and here the linear factor model will be the first tool to which the analyst turns. Second, it may be the case that unsystematic biases and model misfits that are nevertheless large enough to be evident at the item level substantially cancel each other when aggregates neglecting the item-level structure are analyzed. This appears to have happened in the transition from item to test level in our real-data example. Whereas Figure 2 shows a somewhat worse fit of the unidimensional IRT model in the focal group, the test-level fit indices suggest that the single-factor model fits the data from both groups extremely well. In a more extreme situation of this kind, it may be preferable to work at the level of testlets or whole tests if no clear conclusions emerge from the item-level analysis (Wainer, Sireci, & Thissen, 1991).

The problem of quantifying measurement bias is both practically important and didactically useful because of the many points where it directs attention to the unity of linear factor analysis and IRT. We expect that further consideration of the subtle differences between these two variants of the same unified model will shed light on other psychometric problems, some of a more foundational character (Guttman, 1955; McDonald, 2003).

Authors' note

The first two authors contributed equally to this work. Correspondence concerning this article should be addressed to Minji K. Lee, Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN 55902. Email: lee.minji@mayo.edu

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (5th ed.). Washington, DC: American Education Research Association.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Chicester, UK: Wiley.

Butterworth, B. (2010). Foundational numerical capacities and the origins of dyscalculia. *Trends in Cognitive Sciences*, *14*, 534–541. doi:10.1016/j.tics.2010.09.007

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57. doi:10 . 1007 / s11336-009-9136-x

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, *10*, 1–19. doi:10.1007/BF02289789

Chalmers, R. P. (2012). `mirt`: A multidimensional item reponse theory package for the `R` environment. *Journal of Statistical Software*, *48*.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32. doi:10 . 1007 / BF02291477

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.

Fraser, C. & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267–269. doi:10 . 1207 / s15327906mbr2302_9

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical*

Psychology, *8*, 65–81. doi:10.1111/j.2044-8317.1955.tb00321.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Holland, P. W. & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kamata, A. & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136–153. doi:10.1080/10705510701758406

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Lee, J. J. (2012a). Common factors and causal networks. *European Journal of Personality*, *26*, 441–442. doi:10.1002/per.1873

Lee, J. J. (2012b). Correlation and causation in the study of personality (with discussion). *European Journal of Personality*, *26*, 372–412. doi:10.1002/per.1863

Lee, J. J. & Lee, M. K. (2016). An overview of the Normal Ogive Harmonic Analysis Robust Method (NOHARM) approach to item response theory. *Quantitative Methods for Psychology*, *12*, 1–8.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519–537. doi:10.1177/0146621608329504

Lewis, C. (1993). A note on the value of including the studied item in the total score when analyzing test items for DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 317–319). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*, 543–566. doi:10.1016/S0160-2896(03)00051-5

Maydeu-Olivares, A. & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020. doi:10.1198/016214504000002069

McDonald, R. P. (1967). *Nonlinear factor analysis*. Richmond, VA: Psychometric Corporation.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117. doi:10.1111/j.2044-8317.1981.tb00621.x

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*, 212–230.

McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, *5*, 675–686. doi:10.1177/1745691610388766

McDonald, R. P. (2013). Modern test theory. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology vol. 1* (pp. 118–143). New York, NY: Oxford University Press.

McDonald, R. P. & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23–40. doi:10.1207/s15327906mbr3001_2

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. doi:10.1007/BF02294825

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473. doi:10.1007/s11336-007-9039-7

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560. doi:10.1007/BF02293813

Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*, 133–142. doi:10.3102/10769986010002133

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Erlbaum.

Penfield, R. D. & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics vol. 26: Psychometrics* (pp. 125–167). Amsterdam, The Netherlands: Elsevier. doi:10.1016/S0169-7161(06)26005-X

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Sinharay, S. & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*, 23–35. doi:10.1111/emip.12024

Steinberg, L. & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response the-

ory to analyze differential item functioning. *Psychological Methods*, *11*, 402–415. doi:10.1037/1082-989X.11.4.402

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210. doi:10 . 1177 / 014662168300700208

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325. doi:10.1007/BF02295289

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi:10.1007/BF02294363

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197–219.

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a Negative Emotionality scale. *Journal of Personality*, *64*, 545–576. doi:10.1111/j.1467-6494.1996.tb00521.x

Citation

Lee, M. K., Lee, J. J., Wells, C. S. & Sireci, S. G. (2016) A unified factor-analytic approach to the detection of item and test bias: Illustration with the effect of providing calculators to students with dyscalculia. *The Quantitative Methods for Psychology*, *12*(*1*), 9-29.

Received: 07/03/2015 ~ Accepted: 05/08/2015

Tables 2 to 5 follows on next page

Figure 4 ■ Each panel displays both the reference (no calculator) and focal (calculator) item characteristic curves, which give the probability of the correct response as a function of the common factor (θ). Green (light shading) corresponds to the reference group, whereas blue (dark shading) corresponding to the focal group. The item parameters are listed in Table 2. The content strands are Algebra (items 1 to 11), Number Sense (12 to 21), Measurement (22 to 24), Geometry (25 to 27), and Statistics (28 to 34).

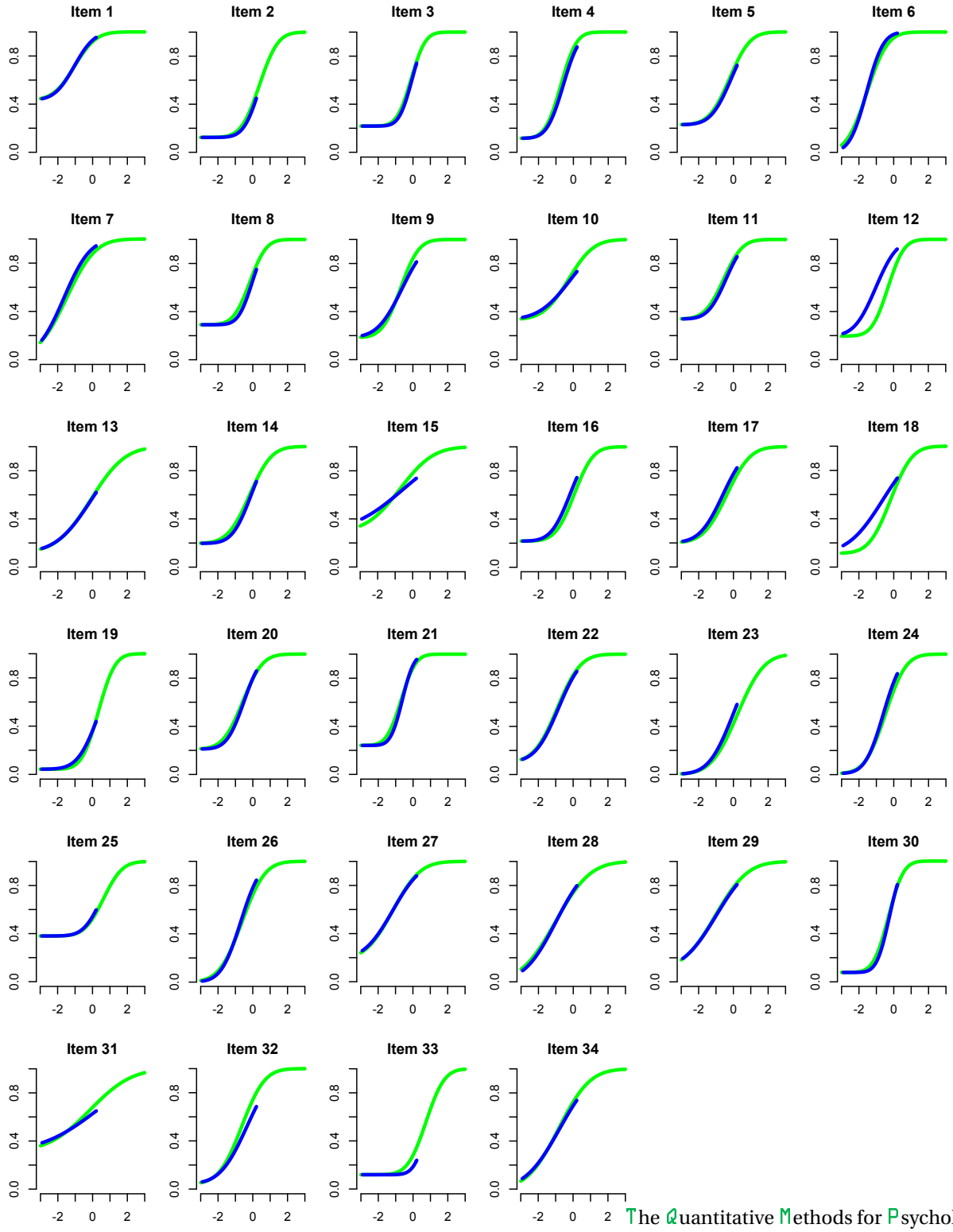


Figure 5 ■ In each panel the line of zero intercept and unit slope is superimposed. The parameters of the focal group (calculator) have been rescaled to the origin and metric of the reference group (no calculator). The coordinates of each point can be found in Table 5. (a) The scatterplot of the intercepts (μ) in the two groups. (b) The corresponding scatterplot of the factor loadings (λ). G, Geometry; M, Measurement; S, Statistics; NS, Number Sense; A, Algebra.

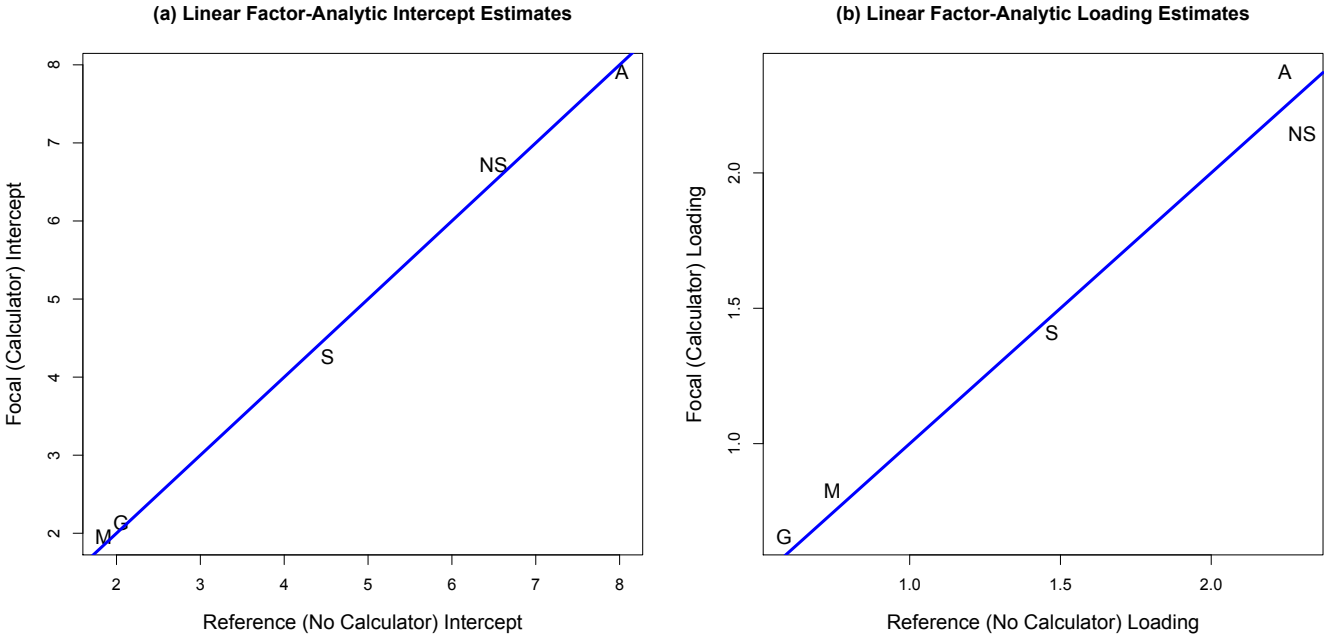


Figure 6 ■ Each panel displays both the reference (no calculator) and focal (calculator) test characteristic curves, which give the expected number of correct responses as a function of the common factor (θ). The dashed curves are the sums of the relevant nonlinear ICCs. Green (light shading) corresponds to the reference group, whereas blue (dark shading) corresponding to the focal group. The linear factor-analytic parameters characterizing the solid lines are listed in Table 5.