A Selective Intellectual History of Differential Item Functioning Analysis in Item Response Theory and Factorial Invariance in Factor Analysis

David Thissen
The University of North Carolina at Chapel Hill

Abstract

The concept of *factorial invariance* has evolved over the past century; it began in the 1930s as a criterion for the usefulness of the multiple factor model, but by the turn of the twenty-first century it had evolved into a form of analysis supporting the validity of inferences about group differences. In the latter form, factorial invariance becomes isomorphic with the lack of *differential item functioning* (DIF) in the literature of item response theory (IRT). This essay traces the evolution of factorial invariance through the twentieth century, with special attention to its vocabulary and elaboration. Then it follows the development of DIF analysis for the past fifty years, concluding that the two kinds of analysis differ in purpose, and therefore details of procedure, but not in their underlying conceptions. Implications are drawn about what each tradition can absorb from the other.

Keywords: factorial invariance, factor analysis, differential item functioning, DIF, item response theory

Introduction

Analyses of *factorial invariance* in factor analysis and *differential item functioning* (DIF) in item response theory (IRT) have separate literatures reflecting their historical development over the past century and half-century, respectively. Meredith (1993) drew together the ideas of factorial invariance with the more recent development of DIF under a superordinate heading of *measurement invariance*, defined as invariance of the distribution of the observed variables, given latent variables, over samples or populations. Thissen (2017) pointed out that while it is established that factorial invariance and a lack of DIF are essentially the same conceptually, in the factor analysis literature invariance is usually examined by blocks of parameters, with associated labels like *metric* and *scalar* invariance, while DIF analysis proceeds item by item, or variable by variable, grouping parameters

Correspondence concerning this article should be addressed to David Thissen, Department of Psychology and Neuroscience, The University of North Carolina at Chapel Hill, 235 E. Cameron Avenue, Chapel Hill, NC, 27599, dthissen@email.unc.edu. Thanks to Michael Edwards, Peter Halpin, Yang Liu, and Anne Thissen-Roe for helpful comments on an earlier version of this ms. Any errors that remain are, of course, my own.

of different kinds. A purpose of this essay is to illuminate the reasons the two traditions are the same yet very different. The difference has its source in the very different questions analyses in the two traditions sought to answer; the sameness arises with the integration of latent variable models within unifying perspectives. We begin with the older tradition, that of factorial invariance.

Factorial Invariance

The Origins of the Idea of Factorial Invariance

Consideration of factorial invariance is essentially as old as multiple factor analysis itself. L. L. Thurstone (1935, p. 55) wrote that "*It is a fundamental criterion for a valid method of isolating primary abilities that the weights of the primary abilities for a test must remain invariant when it is moved from one test battery to another test battery*" (emphasis in the original).[1] There is a qualification later on p. 55: "This criterion assumes that the several test batteries are given to the same population. The primary abilities that define a test in one population should be identical with the primary abilities which define it in a second population." The text is a little confusing about the relationship between what Thurstone was thinking about invariance and distinct populations, but it is clear that his primary focus was the idea that the relation between observed variables and the underlying factors should be the same if variables are embedded within different sets. L. L. Thurstone (1935, p. 120) subsequently nearly repeats himself, again with emphasized text, "One of the fundamental requirements of a successful factorial method is that *the factorial description of a trait must remain invariant when the trait is moved from one battery to another which involves the same common factors or abilities.*"[2] Two points are clear: One is that the origin of the idea of factorial invariance was to buttress the validity of the factor analysis model as a representation of the data and an underlying psychological structure. The second point is that Thurstone looked for this invariance across different combinations of observed variables.

L. L. Thurstone (1947, pp. 360-376) had an entire chapter on "factorial invariance" in what amounted to the second edition of his book on factor analysis. On p. 361, Thurstone quotes himself from the 1935 volume "It is a fundamental criterion ..." L. L. Thurstone (1947, p. 360) waves off invariance across populations (emphasis in the original): "*factor loadings cannot be expected to be invariant from one population to a different population*. Any criterion of invariance in factor analysis assumes that it is applied to analyses on the same population or to equivalent populations."[3] L. L. Thurstone (1947, p. 363) continues to concentrate on invariance across battery composition: "The factorial composition of a test in a set of primary factors that have been found in a complete and over-determined simple structure remains invariant when the test is moved to another battery involving the same common factors and in which there are enough tests to make the simple structure complete and overdetermined."

---

[1] It happens that the source material for this essay includes the copy of *The Vectors of Mind* that Thurstone signed and sent to Harold Hotelling upon publication of the book; that volume is now in the library of the L.L. Thurstone Psychometric Laboratory. The book includes (presumably Hotelling's) pencilled hand-written marginalia. Hotelling's marginal comment on these lines is "Is this possible?" That suggests that invariance will be a topic of research going forward.

[2] Hotelling's marginal comment on p. 120 is "impossible." Possibly Hotelling was thinking of his method of principal components analysis (Hotelling, 1933), in which any change in the set of variables analyzed would change all the weights.

[3] Thurstone footnotes this with "First ed., p. 55", referring to the lines quoted above from his 1935 *The Vectors of Mind*.

While Thurstone concentrated on invariance over the composition of test batteries, Thomson (1950), elaborating on work by Thomson and Ledermann (1939), was more concerned with invariance over populations; they expressed populations mathematically as *selection* of subsets of a super population. Thomson (1950, p. 292*ff*) conceded that univariate selection would yield, as L. L. Thurstone (1947) had previously claimed, invariant factor structures, but then Thomson (1950, p, 294) claimed that multivariate selection may yield different results. It was a rare difference of opinion between Thurstone and Thomson,[4] that Thurstone thought more about test-battery composition and Thomson was more concerned with population differences, which Thurstone believed to be unimportant. Both, however, considered invariance a criterion for the validity of underlying factors.

The Evolution of the Idea of Factorial Invariance: The Mid-twentieth Century

L. L. Thurstone (1947, pp. 363*ff*) has a section on "types of factorial invariance" which states that (p. 363) "The simplest form of factorial invariance concerns the *metric invariance* of the factor matrix solution." He clarified that this means "the numerical values of the factor matrix solution are invariant", and discussed the need for the same reference axes for this to be observed. In the 1940s, the only meaning of the term *factor analysis* was what is now called *exploratory factor analysis*, and for all practical purposes, the only parameters were the factor loadings in the "factor matrix solution." L. L. Thurstone (1947, pp. 364) wrote "A different type of factorial invariance appears in relation to the selection of subjects to whom a test battery is given. This is *configurational invariance*." Thurstone elaborated on p. 365 that "*configurational invariance* is a far more important consideration in factor analysis than the invariance of the factor loadings." Thurstone took it as given that the numerical values of the loadings, and the correlations among the factors, would differ somewhat between populations; he wanted the latent variables, as defined by their salient indicators, to be the same.

In the 1950s, others worked out methods of factor rotation to congruence to show the extent to which factor solutions obtained from different populations were invariant; rotational methods by Ahmavaara (1954) and Kaiser (1958) became available for use, and focus shifted from the question of invariance over battery composition to invariance between groups. And at some point Thurstone's term *configurational* lost some of its syllables: Sharpe and Peterson (1971, p.260) used the shorter term *configural invariance* to mean "factor identities are fundamentally the same"; Guilford, Guilford, and Hoepfner (1971, p. 40) also used "configural invariance" in a similar sense. It is not clear when or how between 1947 and 1971 the word was simplified.

A psychometric literature grew up on how selection could produce (sub-)populations with different covariance or factor structures; a review of that would be an excursion onto a sidetrack at this point. Meredith (1964) assembled and expanded results on invariance of loadings over selected populations, and the consequent invariance of the coefficients for the computation of factor score estimates.

A major step in the history of factorial invariance took place when Karl Jöreskog developed what has come to be called *confirmatory factor analysis* (CFA). Jöreskog (1971, p. 409) described for the first time multi-group factor analysis, writing that the "method is capable of dealing with any degree of invariance, from the one extreme, where nothing is invariant, to the other extreme,

[4] T. G. Thurstone (1980) remarked in a presentation on the history of psychometrics that "Godfrey Thomson was Leon's best friend" ("Leon" being the way Thelma Thurstone referred to Louis Leon Thurstone).

where everything is invariant." Jöreskog used no explicit vocabulary to indicate whether particular parameters (loadings, unique variances, factor correlations) were invariant. The algorithms and software he described could test the equality across groups of any combinations of parameters, using the likelihood-ratio (LR) test statistics that came along with maximum likelihood estimation for the factor model.

Jöreskog (1971) did suggest a sequence of hypothesis tests, beginning with a test that the covariance matrices are (all) equal, then whether the same number of (completely unrestricted) factors could be used for all populations, then (on p. 423) an "invariant unspecified (unrestricted) factor pattern" with each factor identified by one of the observed variables. This latter is not as restrictive as Thurstone's configurational invariance, which was for simple structure.[5] Jöreskog's (1972, p. 424) next hypothesis was basically Thurstone's metric invariance, wherein the loadings were equal across groups for an independent clusters solution for a three-factor example in "a non-overlapping group structure, where the first three tests are loaded on the first factor only, the next three tests on the second factor only and the last three tests on the third factor only." Jöreskog's next test was the hypothesis that the unique variances were also equal across groups, and then a test of the equality of the factor covariance matrices. Following his proposed sequence of model tests with an empirical example, Jöreskog (1971, p. 425) arrived at "two alternative descriptions of the data. One is that the whole factor structure is invariant over populations with a three-factor solution of a fairly complex form. The other is to represent the tests in each population by three factors of a particularly simple form, but these factors have different variance-covariance matrices in the different populations."

An immediate next step followed when Sörbom (1974) expanded upon Jöreskog's work with the estimation of the group means and variance covariance matrices, thus separating group differences in factor means, variances, and covariances from invariance (or the lack of it) in the parameters relating the variables to the factors (the loadings). Sörbom (1974, p. 230) wrote an elaborated factor model as

$$\mathbf{x}_g = \boldsymbol{\mu} + \boldsymbol{\Lambda}_g \boldsymbol{\xi}_g + \boldsymbol{\epsilon}_g \ , \tag{1}$$

where $\mathbf{x}_g$ is a vector of observations from group g, $\boldsymbol{\mu}$ is a vector representing the "origin of the measurements", $\boldsymbol{\Lambda}_g$ is the matrix of factor loadings, $\boldsymbol{\xi}_g$ is the vector of values of the common factors, and $\boldsymbol{\epsilon}_g$ is the random vector of unique factors. Additional parameters were $\boldsymbol{\theta}_g$ and $\boldsymbol{\Phi}_g$, the vector of means and the covariance matrix of $\boldsymbol{\xi}_g$, and $\boldsymbol{\Psi}^2_g$, a diagonal matrix of the variances of the unique factors.

Sörbom (1974, p. 229-230) wrote about "factorial invariance in various degrees", distinguishing among

"1. *Parameters differentiating the measurements*. The vector $\boldsymbol{\mu}$ indicates the level of the measurements, i.e. given $\boldsymbol{\xi}_g = \mathbf{0}$, the expected values of the measurements equal $\boldsymbol{\mu}$. The ith row of $\boldsymbol{\Lambda}_g$, gives the regression of x_i on the common factors and the ith diagonal element of $\boldsymbol{\Psi}^2_g$, is the corresponding residual variance.

"2. *Parameters describing the factor space*. The matrix $\boldsymbol{\Phi}_g$, is the variance-covariance

[5] L. L. Thurstone (1947, p. 181) wrote that "When a factor matrix reveals one or more zeros in each row, we can infer that each of the tests does not involve all the common factors that are required to account for the intercorrelations of the battery as a whole. This is the principal characteristic of simple structure." A commonly hypothesized special case of simple structure is an independent clusters solution, in which each test has a non-zero loading on only one of the several factors.

matrix of the common factors and $\boldsymbol{\theta}$, is their mean vector."[6]

Sörbom (1974, p. 234) followed Meredith (1964) and Joreskog (1971) by using data from Holzinger and Swineford (1939) for his example in which he computed estimates of the factor means for four groups for one of Jöreskog's invariance models.

Jöreskog and Sörbom's models and methods were implemented in a series of increasingly general computer programs culminating in LISREL (Jöreskog & Van Thillo, 1972; Jöreskog & Sörbom, 1974), which became for a time the *de facto* standard software used to investigate factorial invariance.

The Modern Era: The Vocabulary of Factorial Invariance

In the seminal article by Meredith (1993, pp. 527-528), "measurement invariance" is "$F(x|w,v) = F(x|w)$ for all (x,w,v)" (x is observed, w is latent, v is group selection); he attributes this definition to Mellenbergh (1989).[7] With "$F(\cdot)$" defined as "the (cumulative) distribution function of the argument" this covers all models for categorical or continuous data. On p. 530 of the same article "weak measurement invariance" is "$\mathbb{E}(x|w,v) = \mathbb{E}(x|w)$" and "$\boldsymbol{\Sigma}(x|w,v) = \boldsymbol{\Sigma}(x|w)$."

Meredith (1993) used somewhat different notation for the factor analysis model than equation 1 above from Sörbom's (1974) article:

$$\mathbf{x} = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z} + \mathbf{u} \ , \tag{2}$$

where $\mathbf{x}$ is a vector of observations, $\boldsymbol{\alpha}$ represents what are now most often called the intercept parameters, $\boldsymbol{\Lambda}$ is the matrix of factor loadings, $\mathbf{z}$ is the vector of values of the common factors, and $\mathbf{u}$ is the random vector of unique factors plus measurement error. This re-notation permits $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to be used for the mean and covariance matrix for the observations. The vector of means and the covariance matrix of the factors are $\boldsymbol{\xi}$ and $\boldsymbol{\Phi}$. The covariance matrix of $\mathbf{u}$ is Ψ.

Writing about factorial invariance, and using the subscript s to indicate the group or population for parameters that differ between them,[8] Meredith (1993, pp. 527-528) says $\mathbf{X}$ is "strongly factorially invariant" if

$$\boldsymbol{\mu}_s = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\xi}_s \tag{3}$$

and

$$\boldsymbol{\Sigma}_s = \boldsymbol{\Lambda}\boldsymbol{\Phi}_s\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_s \ , \tag{4}$$

and strictly factorial invariant if equation 3 holds and

$$\boldsymbol{\Sigma}_s = \boldsymbol{\Lambda}\boldsymbol{\Phi}_s\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \ . \tag{5}$$

Meredith (1993, p. 538) then summarizes the relationship between factorial invariance and the new concept of measurement invariance as follows: "*If X is strictly factorial invariant with respect to selection on V, X is almost certainly weakly measurement invariant with respect to V.*"

Meredith (1993, p. 541) offered a list of possible outcomes of the analysis of factorial invariance in which successive steps removed differences specified in previous steps one by one,

[6] As will become clear in a subsequent section, the distinction between the first category of parameters associated with the variables and the second category of parameters of the population distributions is crucial in IRT-based DIF analysis.

[7] Mellenbergh (1989) did not use the phrase *measurement invariance*; the definition was for a lack of *item bias*, an earlier term for DIF.

[8] Presumably the s subscript alludes to the tradition that considers multiple groups or populations to be selections from some super population.

by adding equality constraints: "Consider simultaneous factor model fitting to disjoint populations which, we insist, must involve modeling means as well as dispersion matrices. The following cases can arise.

> I. Different factor pattern matrices and different means and variances of the unique (specific plus error) factors over groups.
>
> II. Different means and variances of the unique factors over groups.
>
> III. Strong factorial invariance, i.e., different unique factor variances over groups.
>
> IV. Strict factorial invariance."

Note that cases I and II had no names.

Table 1 lists some of the vocabulary used to describe analyses of factorial invariance that has grown from those seeds. When the only factor analysis was what is now called exploratory factor analysis, L. L. Thurstone (1947) had distinguished between *configurational* and *metric* invariance; the former referred to the same simple structure across groups, and the latter to the same loadings.

Widaman and Reise (1997) discuss *configural invariance* (p. 292), attributing the term to Horn, McArdle, and Mason (1983) although it is certainly derivative of Thurstone's (1947) *configurational*, with the term shortened to *configural* by the early 1970s. Widaman and Reise (1997, p. 293) wrote "Meredith (1993) distinguished several forms of factorial invariance, forms that he termed *weak*, *strong*, and *strict* factorial invariance.[9] For Widaman and Reise (p. 293), weak factorial invariance has equal loading matrices across groups. Then Widaman and Reise (p. 294) have strong factorial invariance, constraining the loadings and intercepts of the measured variables; that corresponds with Meredith's definition (p. 534) of "strongly factorial invariant." To obtain strict factorial invariance, Widaman and Reise (p. 295) add the constraint that the unique factor variances are equal across groups.[10] Widaman and Reise go on to say "All these are forms of metric invariance" which is no doubt true in a semantic sense, but that deviates from the more specific use of the term *metric* invariance by Thurstone (1947).

Going forward in time, the terms *weak* and *strong* have been used with different meanings, sometimes both meaning the same model for different authors. Setting those terms aside, there has been convergence on the use of *configural* for the same factor pattern, *metric* for equal loadings across groups, *scalar* for equal intercepts as well, *strict* if all parameters of a CFA model are equal across groups.

There have been uses other than Meredith's (1993) of the term *measurement invariance* in the factor analysis literature. Horn and McArdle (1992, p. 117) used it with a verbal definition, "the same attribute is measured in the individuals within groups." Writing in a similar vein, Reise, Widaman, and Pugh (1993, p. 552) wrote "To compare groups of individuals with regard to their level on a trait, or to investigate whether trait-level scores have differential correlates across groups, one must assume that the numerical values under consideration are on the same measurement scale (Drasgow, 1984, 1987). That is, one must assume that the test has 'measurement invariance' across groups." Reise et al. (1993, p. 552) did not cite Meredith, and it seems their use of the term "measurement invariance" was independent, although they also used it to encompass both the analysis

[9] Meredith (1993) did not use the modifier *weak* for factorial invariance, except by implication of the use of *strong* that there must be some meaning to *weak*.

[10] All these cases (configural, weak by implication after Meredith's use of "strong", strong, and strict leave the group factor means and variance-covariance matrices unconstrained, as we shall see would be the case in IRT-based DIF analysis.

Source	Same zeroes in $\mathbf{\Lambda}$	Equal $\mathbf{\Lambda}$	Also Equal $\boldsymbol{\alpha}$	Also Equal $\mathbf{\Psi}$
L. L. Thurstone (1947, pp. 363*ff*)	*configurational* invariance	*metric* invariance		
Horn and McArdle (1992, pp. 123*ff*)	*configural* invariance	*metric* invariance		
Meredith (1993, p. 534)		(unnamed case II)	*strong* factorial invariance	*strict* factorial invariance
Widaman and Reise (1997, pp. 292-295)	*configural* invariance	*weak* factorial invariance	*strong* factorial invariance	*strict* factorial invariance
		"All these are forms of metric invariance"		
Steenkamp and Baumgartner (1998, pp. 80-81)	*configural* invariance	*metric* invariance	*scalar* invariance	*error variance* invariance
Vandenberg and Lance (2000, pp. 12-13)	*configural* or *weak* invariance	*metric* or *strong* invariance	*scalar* invariance	"invariant uniquenesses"
Millsap and Meredith (2007, p. 133-134)	*configural* invariance	*metric* invariance or *weak* factorial invariance	*scalar* invariance or *strong* factorial invariance	*strict* factorial invariance
Putnick and Bornstein (2016, p. 74)	*configural* invariance	*metric* invariance or *weak* factorial invariance	*scalar* invariance or *strong* factorial invariance	*residual* invariance or *strict* factorial invariance

Table 1

Vocabulary used by various authors for some patterns of invariance, organized by columns representing model matrices in equations 3-5.

of factorial invariance and IRT-based DIF analysis. That article emphasized the consideration of partial invariance; that brings the analysis of factorial invariance closer to DIF analysis.

Vandenberg and Lance (2000) used structural equations modeling as the basis of their use of the term *measurement invariance*, distinguishing between that (subsuming the tests of "configural", "metric", "scalar", and "uniquenesses" invariance), and "structural invariance"; the latter included the factor variances, covariances, and means. So Vandenberg and Lance (2000, pp. 12-13) added tests of equality across groups of factor variances, covariances, and means as three more steps/tests after the by then canonical tests of configural, metric, scalar, and strict invariance. This separation actually goes back to Sörbom's original work with group means and covariance matrices, and language commonly used for structural equations models: The parts of the model that link latent variables to observed variables are called the *measurement model* and the parts that link latent variables to each other (either through correlation or regression) make up the *structural model*.

Partial Factorial Invariance

Putnick and Bornstein (2016) have a section on *partial invariance*, describing models in which some, but not all, parameters in a block are invariant across groups. In subsequent sections we will see that partial invariance is an essential part of the conceptualization of DIF. It has rarely been mentioned in the factor analysis literature, but for item response theorists partial invariance models are the essence of DIF analysis, intended to identify and then remove the observed variables (items) that are not invariant; more on that in a subsequent section. According to Van De Schoot, Schmidt, De Beuckelaer, Lek, and Zondervan-Zwijnenburg (2015, p. 1), Byrne, Shavelson, and Muthén (1989) are the origin of the term "partial measurement invariance"; Byrne et al. (1989) did not use the kind of vocabulary summarized in Table 1, but they did discuss the possibility of partial invariance.

Reise et al. (1993, p. 555) considered "full measurement invariance" to hold if the factor loading matrices are equal across groups, and "partial measurement invariance" (p. 556) if some, but not all, of the loadings are equal across groups. Reise et al. (1993) compared CFA results obtained with that conception of full and partial "measurement invariance" with IRT-based DIF analysis results, for which they defined "full measurement invariance" (p. 561) as *all* item parameters (slopes and thresholds/intercepts) equal across groups, with only the group latent means different, and "partial measurement invariance" (p. 561) as models with some items' parameters equal across groups[11] and some not equal. The presentation by Reise et al. (1993) was not completely parallel between the CFA and IRT analyses, because the intercept parameters and means were omitted from consideration in the analyses of factorial invariance.

Widaman and Reise (1997, p. 299) also wrote briefly about partial factorial invariance; they said that "many experts on structural modeling argue that partial metric factorial invariance models are not viable models for demonstrating that true ARF-invariant latent variables are identified."[12] Unlike Widaman and Reise (1997, p. 299), Steenkamp and Baumgartner (1998, pp. 80-81) mention partial invariance without negative connotation. Vandenberg and Lance (2000) also wrote about partial invariance.

[11] The items with equal parameters across groups are those that form the "anchor."

[12] Widaman and Reise (1997, p. 291) define "ARF invariance" as "invariance under appropriate rescaling factors", meaning "essential relations among the group means are invariant under the different rescalings."

Contemporary Syntheses

Putnick and Bornstein (2016, p. 73) describe a "ladder" of factorial invariance derived from "landmark papers, Widaman and Reiss (1997) and Vandenberg and Lance (2000), [that] synthesized the measurement invariance literature, delineated the ladder-like approach to measurement invariance testing, and provided researchers with step-by-step guides to conducting invariance tests." Their ladder was *configural invariance*, *metric invariance* (equal loading matrices), *scalar invariance* meaning the intercepts are also equal; then *residual invariance*. This current "state of the art" (Putnick & Bornstein, 2016, p. 73) is largely due to the synthesis by Millsap and Meredith (2007, p. 133-134):

> If configural invariance is tenable, the next step is to constrain the factor pattern matrices to invariance . . . , a condition known as *weak factorial invariance* (Widaman & Reise, 1997) or metric invariance (Horn & McArdle, 1992; L. L. Thurstone, 1947). Weak factorial invariance implies that any systematic group differences in the covariances among the measured variables are due to the common factors, rather than other sources of association. If weak factorial invariance is retained, the measurement intercepts are constrained to invariance next . . . , yielding strong factorial invariance (Meredith, 1993) or scalar invariance (Steenkamp & Baumgartner, 1998). Strong factorial invariance implies that any systematic group differences in either the means or the covariances among the measured variables are due to the common factors. The final step imposes invariance on the unique factor variances . . . , leading to strict factorial invariance (Meredith, 1993). Strict factorial invariance implies that any systematic group differences in the means, variances, or covariances for the measured variables are due to the common factors, rather than group differences in factor structure. For example, systematic group differences in the means on the measured variables are due to differences in the factor means. Strict factorial invariance is useful if present because it clarifies the interpretation of any group comparisons on observed means or covariance structures.

Recent Extensions

Van De Schoot et al. (2015, p. 1) wrote "As proposed by Mellenbergh (1989), "measurement invariance" (MI) requires that the association between the items (or test scores) and the latent factors (or latent traits) of individuals should not depend on group membership or measurement occasion (i.e., time). In other words, if item scores are (approximately) multivariate normally distributed, conditional on the latent factor scores, the expected values, the covariances between items, and the unexplained variance unrelated to the latent factors should be equal across groups." Van De Schoot et al. (2015, p. 1) subsequently describe

> A relatively new research avenue in the MI literature deals with the use of Bayesian structural equation models (BSEM) to relax strict forms of MI (see Muthén and Asparouhov, 2012). In particular, exact zero constraints on the cross-group differences between all relevant measurement parameters (e.g., factor loadings and item intercepts) are substituted by "approximate" zero constraints. Instead of forcing item intercepts to be exactly equal across groups, a substantive prior distribution (around zero) is used to bring the parameters closer to one another, while allowing for some "wiggle room."

> If there are many small differences between the groups in terms of intercepts or factor loadings, approximate MI seeks a balance between adherence to the requirements of MI, making comparisons possible, and obtaining a well-fitting model (i.e., a model that is more realistic given the data at hand). When the classical MI tests do not hold given the data, approximate MI represents a promising (and more realistic) alternative; the cross-group differences between all relevant measurement parameters are "hopefully" close enough to zero to allow making meaningful latent factor mean comparisons.[13]

Thus, by the first part of the twenty-first century, the analysis of factorial invariance has evolved from a test of the validity of the factor model (which it was for Thurstone and Thomson) into analysis supporting the validity of inferences about group differences. That brings the purpose of the analysis of factorial invariance closer to that of the analysis of differential item functioning in the IRT literature.

Differential Item Functioning in IRT

While DIF analysis within an IRT framework and the analysis of factorial invariance are now considered in some senses the same within a superordinate concept of measurement invariance, the two forms of analysis have very different origins and purposes. The following sections trace the development of DIF analysis to its convergence with factorial invariance, and then provide reflections on that convergence.

The Origins of DIF Analysis

In his seminal chapter on latent structure analysis, Lazarsfeld (1950, p. 367) wrote "We shall now call a *pure test* of a continuum x an aggregate of items which has the following properties: *All interrelationships between the items should be accounted for by the way in which each item alone is related to the latent continuum.*" Then later on that same page he continued "we would want our analysis to clarify what is meant by group factors in itemized tests, what is meant by biased questions, and so on. It will turn out that all of that can be achieved by comparing an actual aggregate of items with the special case of a pure test."

It would be more than twenty-five years before what was to become IRT would identify "biased questions," but that time would come. The incorporation of the study of "biased questions" into IRT can be traced to a conference presentation by Lord (1977); in his subsequent expansion of that presentation, Lord (1980, p. 212) wrote with characteristic succinctness "It is frequently found that certain disadvantaged groups score poorly on certain published cognitive tests. This raised the question of whether the test items may be unfair or biased against these groups. . . . If . . . an item has a different item response function for one group than for another, it is clear that item is biased. If the bias is substantial, the item should be omitted from the test."

So DIF analysis originated in educational measurement, with the goal of detecting and removing "biased" items, to reduce overall unfairness against some demographic group(s). It has become standard practice to include DIF detection in item analysis for any kind of test construction (Edwards & Edelen, 2009). This is in contrast to the origins of factorial invariance that involved the theoretical question of whether sampling a set of variables, or a subgroup from a population, would affect the factor structure, or more generally the validity of the factor analysis model. A

[13] This idea will reappear subsequently when regularized DIF analysis is discussed.

consequence of this difference in motivation is that, in the beginning, DIF analysis was primarily about the difficulty of test items (on tests of proficiency or achievement; that is analogous to item endorsement rates on personality or social questionnaires, and the variables' intercepts in factor analysis), whereas the beginnings of the study of factorial invariance were exclusively about the relationships of the observed variables to the underlying factors (the loadings). Consideration of factor-analytic intercept parameters, the equivalent of item difficulty, did not arise in conceptions of factorial invariance until the 1970s.

From Item Bias to DIF

Before Lord (1977, 1980) brought the topic within the purview of IRT, item bias had been studied using simpler statistical procedures. Angoff and Ford (1973) plotted the ETS delta [14] associated with the the inverse normal transformation of the percent correct on SAT items for Black and White students, and noted that outliers on that scatterplot probably represented item bias. Lord (1977) objected that the inverse normal transformation of the percent correct was a poor indicator of item difficulty, and suggested the use of the IRT difficulty parameter instead. Angoff and Ford (1973) and Lord (1977, 1980) used data from the SAT, and Black and White examinees as groups. For the running example in this section, data from a spelling test[15] administered to college undergraduates will be used. The item parameters, including the IRT difficulty parameter b, were obtained by fitting the 2PL IRT model to a subset of 76 items of the 100 item spelling test,[16] using data from 1000 examinees. This yields the scatterplot shown in Figure 1, with the difficulty ($b_{Females}$) of the items for female examinees plotted against difficulty (b_{Males}) of the items for male examinees. Items represented by points above and to the left of the diagonal reference line are relatively more difficult for female examines to spell, while items with points below and to the right of that line are more difficult for males. *Difficulty*, here, reflects the probability the word will be spelled incorrectly, and b is a parameter of the two-parameter logistic (2PL) IRT model.

The Two-Parameter Logistic (2PL) IRT Model

The 2PL IRT model (Birnbaum, 1968), describes the probability of a correct item response as a function of item parameters (a_i and b_i for each item i), and θ, the latent variable measured by the test, as follows: The conditional probability, or *trace line*, of a correct or positive response $u = 1$, as a function of the latent variable being measured (θ) is

$$T(u_i = 1|\theta) = \frac{1}{1+\exp[-a_i(\theta - b_i)]} \ . \tag{6}$$

[14] ETS delta is a linear transformation of the standard score scale to have a mean of 13 and a standard deviation of four. It was used by item analysts at Educational Testing Service as a kind of *lingua franca* across tests scored on different scales.

[15] The purpose and content of the spelling test are described by Bock, Zimowski, and Thissen (1997).

[16] As described by Bock et al. (1997), the words on the spelling test were chosen using random sampling from a long list of words in the English language. That created an explicitly item-sampled test for the Bock et al. (1997) study of domain sampling, but it had the property that a number of the items performed poorly: They were too easy or too difficult, or not particularly discriminating, for the college student sample. To better represent the performance of the DIF procedures discussed here, 24 of the 100 original spelling items have been set aside, leaving a 76-item test. Angoff and Ford (1973) and Lord (1977, 1980) used all of the SAT items; the items in their examples were already highly selected for their psychometric properties.

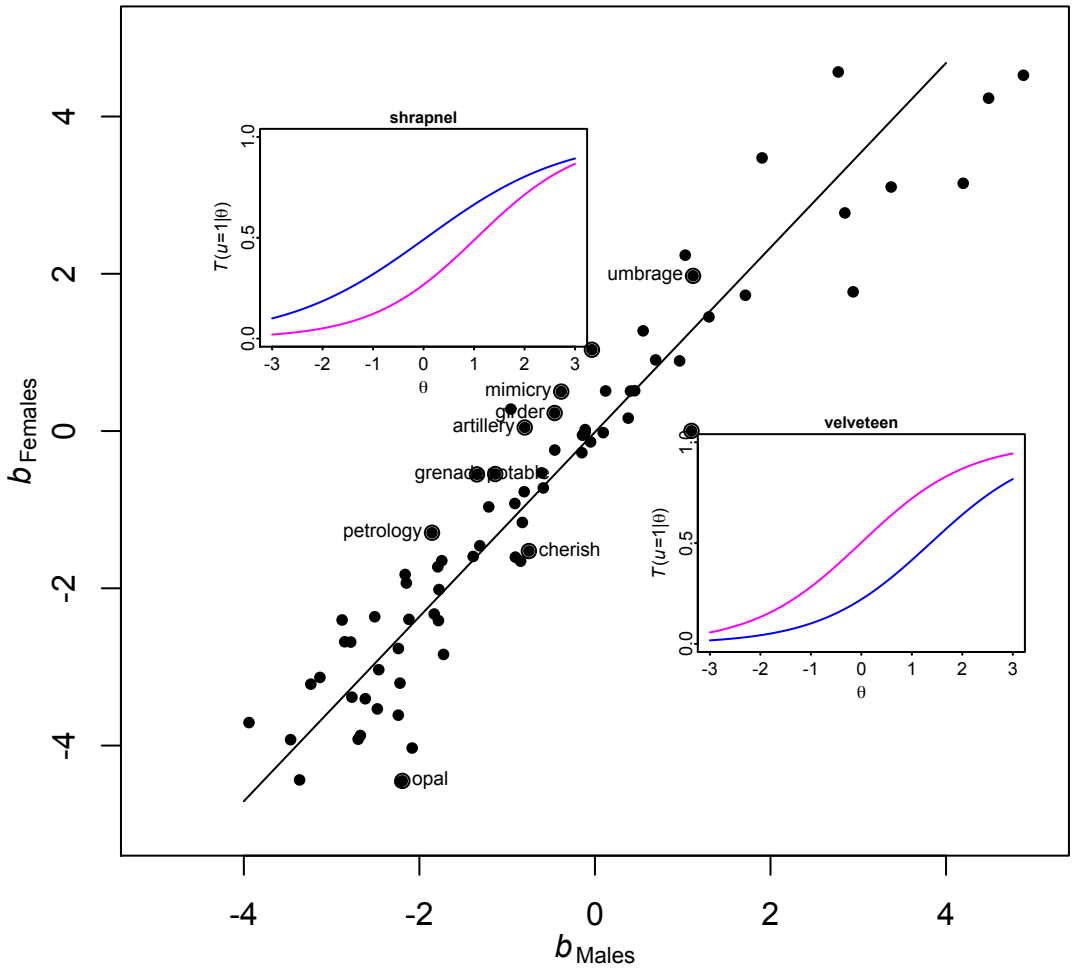

Figure 1

2PL difficulty ($b_{Females}$) parameters of the spelling items for female examinees plotted against difficulty (b_{Males}) parameters of the items for male examinees. Items represented by points above and to the left of the diagonal principal axis reference line are relatively more difficult for female examines to spell, while items with points below and to the right of that line are more difficult for males. The 2PL trace lines for the male (blue) and female (magenta) examinees for spelling the words shrapnel *and* velveteen *are shown in the inset boxes attached to the points for those words, upper left and lower right.*

The thresholds b_i are the values on the θ scale at which a respondent has a 50% chance of responding correctly or positively; the slope a_i is the regression coefficient of the log odds of a correct or positive response regressed on θ.[17] The latent variable θ is usually standardized in one of the groups (in this case the males); that sets the scale for the parameters a and b.

Early Methods of DIF Detection

DIF analysis rests on the following idea: "If ... an item has a different item response function for one group than for another, it is clear that the item is biased" (Lord, 1980, p. 212). Inset panels in Figure 1 show the trace lines (equation 6) for the *reference* group (here, male, blue) and the *focal* group (female, magenta) for spelling the words *shrapnel* (upper left) and *velveteen* (lower right). The differences between the pairs of trace lines are one indicator of the effect size for DIF (Steinberg & Thissen, 1996); another effect size is the difference between the b parameters themselves. The inset graphics in Figure 1 provide meaning for those two points on the scatterplot; points along the diagonal reference line correspond to trace lines that are effectively coincident for the males and females.

In the context of parametric IRT, there is a one to one relation between item response functions (Lord's term for the trace lines $T(u_i = 1|\theta)$) and the item parameters, so the question of whether "an item has a different item response function for one group than for another" is answered with a statistical test of the equality of the item's parameters for one group and those for the other. Such a statistical test has two elements: One is some mechanism to determine, and "correct for," whatever overall differences exist between the groups on the latent variable measured by the test or scale. That is usually done by designating some set of items as the *anchor*, by analogy with the anchor in test linking designs. While the best way to designate the anchor is on some theoretical grounds (Thissen, Steinberg, & Wainer, 1993), that is rarely done in practice, and much of the development of IRT-based DIF analysis involved selection of the anchor item set.

The second element is the statistical test itself. Wald tests and LR tests of the significance of the difference between parameters for the focal and reference groups have been used most often; score (or Lagrange multiplier) tests have also been suggested (e.g., by Schneider, Strobl, Zeileis, and Debelak (2022) and Zimmer, Draxler, and Debelak (2022)).

Lord (1977, 1980) used different approaches to (effectively) select an anchor item set: Lord (1977) used the principal axis of a plot like the one shown in Figure 1 to determine the linear transformation that placed the item parameters for the two groups on the same scale after they had been estimated separately, then he used Wald tests to evaluate the significance of the difference between item parameters for the two groups. After finding that 46 of the 85 items were significantly different at the $\alpha = 0.05$ level of significance, Lord re-estimated the item parameters for the two groups using only the subset of 32 items that were not (preliminarily) different at the $\alpha = 0.15$ level to obtain estimates of θ uncontaminated with any item bias. Then he used those values of θ as fixed to estimate the parameters of all 85 items on the same scale, and again used Wald tests of significance, finding 38 of the 85 items to have parameters significantly different between groups.

In a substantial revision of that original work, Lord (1980, pp. 217*ff*) proposed somewhat different approaches. Instead of separately calibrating the items in the two groups, standardizing θ

[17] All of the methods of DIF analysis to be described in this and subsequent sections can be generalized to other parametric IRT models, with different parameterizations, more parameters, or more response categories (see van der Linden (2021)). However, the central principles remain the same, so the focus here will remain on the parameters of the 2PL model.

in each group to set the units of the scale, the alternative idea was "standardizing on the b": fixing the mean and standard deviation of the b parameters to be zero and one, and counting on that to place the item parameters of both groups on the same scale. Wald tests could be used to test the between group difference of the parameters for each item. Lord (1980, p. 220) then suggested the idea of "purification of the test": If many items were found to be biased in the preliminary analysis, a second analysis could be based on fixed values of θ computed using the items remaining after removing those with significantly different response functions in the preliminary analysis. Re-estimated item parameters based on those fixed values of θ were then tested for between-groups significance. So the "purification" of the test was to (effectively) use as the anchor the subset of items with the most similar parameters between groups.

The conception of IRT in Lord's (1980) volume, assembling and revising his seminal research of the 1970s, did not include reference to a population distribution for the latent variable θ; the same was true of the software LOGIST (Wood, Wingersky, & Lord, 1976) that was used to do the computations. A consequence was that item parameters could only be estimated on a scale defined for a single group at a time, and procedures such as those described above were necessary to put the parameters "on the same scale" (Lord, 1977, p. 25). The slope (A) and the intercept (B) of the linear transformation of the independently-estimated focal group parameters to the reference group scale are the slope and intercept of the reference (principal axis) line illustrated in Figure 1.

However, there is a population distribution for θ for each group in a complete conceptualization of the IRT model; those population distributions are often characterized as normal with mean μ and standard deviation σ. Kolen and Brennan (2004, p. 164) summarize the relationships between the rescaling parameters A and B and the means and standard deviations of θ and b-parameter distributions as follows, with (here) the subscripts R and F referring to reference and focal groups:

$$A = \frac{\sigma(b_F)}{\sigma(b_R)} \ , \tag{7}$$

$$= \frac{\mu(b_F)}{\mu(b_R)} \ , \tag{8}$$

$$= \frac{\sigma(\theta_F)}{\sigma(\theta_R)} \ , \tag{9}$$

$$B = \mu(b_F) - A\mu(b_R) \ , \quad \text{and} \tag{10}$$

$$= \mu(\theta_F) - A\mu(\theta_R) \ . \tag{11}$$

So one way to determine the rescaling coefficients A and B is to determine the slope and intercept of the principal axis in a plot such as Figure 1, perhaps even graphically; Lord (1977) did that. Another way is to use the relations involving the bs in the equations above: Kolen and Brennan (2004, p. 167) refer to using equations 7 and 10 as the *mean/sigma* method (Marco, 1977) and 8 and 10 as the *mean-mean* method (Loyd & Hoover, 1980). From the mid-1980s onward those methods were mostly superseded in both DIF analysis and IRT test-linking by a method proposed by Stocking and Lord (1983), effectively choosing A and B to minimize the squared difference between the test characteristic curves (TCCs—the expected summed score as a function of θ).

However, equations 9 and 11 make it clear that another way to have the item parameters on the same scale is to directly estimate the mean and standard deviation of the distribution of θ for

the focal group relative to that of the reference group along with the item parameters in a process called *concurrent calibration*. If the reference group population distribution has $\mu = 0$ and $\sigma = 1$, the values of B and A are the mean and standard deviation of the focal group. Then there is no need for an explicit linear transformation of the item parameters for one group onto the scale of the other, because in concurrent calibration all of the parameters are already estimated on the same scale.

Lord (1977, 1980) used Wald tests to evaluate the statistical significance of the difference between item parameters for the focal and reference groups; however, those Wald tests were based on estimates of the variance covariance matrix that considered the values of θ to be fixed-and-known in the software LOGIST (Wood et al., 1976). The fact that θ is an unknown random variable in the model made those under-estimates of the variances in the covariance matrix, and inflated the Wald test statistics.

Thissen, Steinberg, and Gerrard (1986) combined concurrent estimation of the group difference and the item parameters with LR tests in an analysis that was parallel with the analysis of factorial invariance, although the 2PL item response model was used. LR tests require only the loglikelihood of the two fitted models being compared; they do not require accurate estimates of the item parameter error covariance matrix, which was a challenge to compute in the 1980s.

Thissen, Steinberg, and Wainer (1988) compared procedures implemented in various software packages at the time, including both Wald and LR tests, in real and simulated data using known item anchor sets. Thissen et al. (1993) followed up with more detailed comparisons of Wald and LR tests using four different modeling-and-software combinations, and some data from the spelling test used for the running example here as well as SAT data. All of these procedures involved concurrent calibration, using some set of anchor items to (effectively) measure the difference between groups, placing the item parameters on the same scale, and then comparing item parameters for the *studied* items between groups. Thissen et al. (1993, p. 102) discussed the value of a "designated anchor": a set of items pre-determined to exhibit no DIF, that effectively determine the difference in distributions of the latent variable between groups. They emphasized that both statistical analysis and expert judgment should be involved in the selection of the anchor items. Williams (1997) described constructing designated anchor sets for a statewide mathematics test, using a few test items alone or test items along with teachers' grades.

Williams's (1993) dissertation, the basis of her 1997 publication, involved hundreds of applications of then-available IRT software, fitting one model at a time, and assembling the results into many LR tests for individual items. In the conclusion of that dissertation, Williams (1993, p. 45) suggested for the future that "What is necessary is computer software which will simultaneously compute the difference between the likelihood ratio statistics for the fully constrained and unconstrained models for all test items under investigation." It would be some years before such software was developed, but it was created in minimal form by Thissen (2001) to obtain simulation results for a symposium on DIF presented at the annual meeting of the National Council on Measurement in Education in 2001. That software include a mode in which it fitted the 2PL (or graded) IRT models to all the items in a test with their parameters constrained equal between two groups; then it freed the constraints on each of the parameters of each item, one item at a time, computing LR test statistics as the difference between the log likelihoods for the constrained and less constrained models. This was done to produce software for the IRT LR DIF procedure that was comparable in effect to that available for observed-score-based DIF procedures.[18] Absent alternatives, the IRTLRDIF (Thissen,

[18]There have been a number of procedures proposed to detect or measure DIF that are not based on statistical tests of the parameters of IRT models. Well known among those are the Mantel-Haenszel procedure (Holland & Thayer, 1988),

2001) software took on a life of its own and was fairly widely used for a decade or more. That software has subsequently been superseded by a different algorithm implemented in the IRTPRO™ (Cai, Thissen, & du Toit, 2011) and flexMIRT® (Houts & Cai, 2020) packages. The mirt software package (Chalmers, 2012) in R, has as one of its options an algorithm with the essential features of IRTLRDIF.

To continue with results obtained with the spelling test, the column of Table 2 labeled $X^2_{LR}(2)$ contains the LR χ^2 tests with 2 $d.f.$ for 18 items with some evidence of DIF relative to the overall item set.[19] All of the items in Table 2 have significant $X^2_{LR}(2)$ values after using the Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995; Edwards & Edelen, 2009) to set the *false discovery rate* at 0.05 across the 76 overall DIF tests. The top 11 items in Table 2 exhibit significant DIF as evaluated by four procedures (two yet to be described); those are the items with labels in Figure 1. The χ^2 tests with 2 $d.f.$ in the column of Table 2 labeled $X^2_{Wald}(2)$ contains the Wald χ^2 tests with 2 $d.f.$ computed with the algorithm described by Woods, Cai, and Wang (2013) and implemented in IRTPRO™ (Cai et al., 2011) and flexMIRT® (Houts & Cai, 2020). One of those, item 54, *tableaux* does not exhibit significant DIF using the Wald test, after correction for multiplicity; that is not unexpected, given results obtained by Woods et al. (2013) that showed that the Wald procedure using the latent mean and variance computed with all the items exhibited a little less power then LR tests using all other items as anchor. There is no indication from any of the analyses of slope DIF for any of the items in the spelling test, so this is exclusively about *b*-DIF, or differential difficulty.

The results in Table 2 were obtained without purification of the anchor. Woods (2009) mentioned three earlier iterative purification procedures by Bolt, Hare, Vitale, and Neman (2004), Edelen, Thissen, Teresi, Kleinman, and Ocepek-Welikson (2006), and Kim and Cohen (1995), then suggested an empirical procedure to select anchor items using only the test responses. Woods (2009) suggested beginning with all-other-items as anchor LR tests, and then selecting the 10-20% of the items with the smallest LR test statistics as the anchor. This assumes that some central (in some sense) set of the items measure the intended construct.

Recent Developments on the Selection of Anchor Items

When All of the Items Exhibit DIF

The heading of this section is a paraphrase of the title of Steinberg's (2016) presentation to the Society for Multivariate Experimental Psychology entitled *When Every Item Exhibits DIF Relative to the Same Anchor: An Illustration of the Interpretation of DIF*. That presentation summarized results described in more detail by Sharp et al. (2019), in which the "items" were nine diagnostic criteria for borderline personality disorder (BPD), rated by a clinical interviewer. In a comparison of the ratings obtained by adolescents and adults, all nine items exhibited significant DIF using the all-items-as-anchor Wald test procedure. That result is surprising, because the DIF detected with that algorithm is relative to the average difference between the groups across those very same nine items. Inspection of the direction of DIF and the item content clarified the result: For a set of four of the criteria ("abandonment fears, unstable relationships, identity disturbance, emptiness") the adults

SIBTEST (Shealy & Stout, 1993), and area between the curves computations (Raju, 1988); there are others. However to maintain parallelism with the analysis of factorial invariance using parametric factor analytic models, the focus here remains solely on DIF procedures based on parametric IRT models.

[19] These values were computed with the mirt software package (Chalmers, 2012) in R.

Item	Spelling Word	$X^2_{LR}(2)$	$X^2_{Wald}(2)$	$X^2_{Robust}(2)$	Wt_{Robust}	$dif_{Regularized}$
39	grenade	47.1	35.6	33.5	0	-0.33
8	shrapnel	46.0	41.3	42.1	0	-0.23
27	girder	28.8	25.1	24.5	0	-0.14
72	artillery	31.9	30.6	30.0	0	-0.13
92	quagmire	18.9	18.1	18.8	0	-0.07
38	petrology	15.3	14.7	15.2	0	-0.05
69	umbrage	12.0	11.6	11.4	0	-0.05
15	mimicry	16.8	16.3	16.7	0	-0.04
53	cherish	18.1	16.7	11.1	0	0.09
47	velveteen	70.6	60.7	52.3	0	0.35
96	opal	59.1	47.3	43.5	0	0.56
		Robust: not (significant) DIF				
35	cheval	12.1	11.3	*10.1*	0	0.07
18	pajamas	12.5	11.7	*9.3*	0.02	0.16
16	diner	12.8	11.7	*8.7*	0.32	0.25
		Regularized: not DIF				
64	potable	13.4	13.5	14.2	0	0
82	invigorating	12.9	11.3	11.1	0	0
		Robust and Regularized: not (significant) DIF				
88	indecent	11.2	10.5	*9.1*	0	0
		LR DIF only (barely)				
54	tableaux	10.6	*8.7*	*5.8*	0	0

Table 2

Results for four DIF detection methods for the 18 (of 76) spelling test items with significant DIF indicated by the LR test using all other items as the anchor. $X^2_{LR}(2)$ *is the LR test statistic value,* $X^2_{Wald}(2)$ *is the Wald test statistic value using the latent variable mean and variance obtained with all items constrained equal,* $X^2_{Robust}(2)$ *is the robust method test statistic,* Wt_{Robust} *is the robust method linking weight, and* $dif_{Regularized}$ *is the regularized* $b-dif$ *estimate. Italized* X^2 *values are not significant after BH control of the false discovery rate.* $Wt_{Robust} = 0$ *or* $dif_{Regularized} \neq 0$ *indicates DIF.*

more often received higher ratings; for the other five criteria ("impulsivity, suicidal behaviors, affective instability, uncontrolled anger, and paranoid ideation") adolescents scored higher (Sharp et al., 2019, p. 1017).

This result is a very different one from the original plan for DIF analysis, which was to find a small number of biased items on a test of educational achievement or proficiency, and remove those. The result converges on the late-twentieth-century intended use of the analysis of factorial invariance, to draw conclusions about whether a scale provides comparable measurement across groups. The item-by-item nature of the DIF analysis converges with the idea of partial factorial invariance. And the report makes it clear that expert judgment is required to make sense of the DIF, which in the case of this example shows that one can conceive of two factors underlying ratings for BPD, one which tends to increase between adolescence and adulthood, and the other which decreases. Further, it makes it clear that DIF results, like the results obtained with exploratory factor analysis, require expert judgment to interpret. The statistical tests alone do not say what is measured by which indicators.

When Many of the Items May Exhibit DIF: Growing the Anchor

DIF analysis has increasingly come to be used to provide evidence supporting the veracity and validity of translations and educational tests and psychological scales. However, some studies have shown that many items on a translated scale may exhibit DIF; for examples see Edelen et al. (2006); Ercikan and Koh (2005); Orlando and Marshall (2002), and Yang et al. (2011). If many items on a scale exhibit DIF, using all items, or all of the other items, as the anchor for DIF tests may yield poor results, because the anchor is highly contaminated. For that reason, to guard against the possibility that many items would exhibit DIF, Y. Chen et al. (2019) and Y. Chen et al. (2020) used a procedure to grow the anchor from a single item in the context of DIF analysis for translations of two sets of clinical psychology scales from English into Mandarin.

The use of one-item anchors has been proposed a number of times: Thissen et al. (1988) mentioned the idea briefly, Rensvold and Cheung (2001); Wang (2004) suggested analyses that rotated through the use of all other individual items on the scale as anchors, one at a time, and Stark, Chernyshenko, and Drasgow (2006) proposed that a single item be chosen by analogy with a single variable that is chosen to set the scale in factor analytic analyses. While Stark et al. (2006) correctly point out that the choice of a single variable to set the scale makes no difference for linear-normal factor analytic results, that is only true for the fit of the model. The choice of a single item to identify the latent variable is very important for the definition of that latent variable, and that is important for item analyses in test theory.

So the variant of the one-item anchor used by Y. Chen et al. (2019) and Y. Chen et al. (2020) differed in two ways from previous suggestions: (1) The (first) anchor item was carefully chosen to be the item that an expert panel considered best represented the trait that was the target of measurement; the experts were also informed of the pattern of DIF results obtained using all items as the anchor. (2) An attempt was made to grow the anchor, by testing all other items for DIF, and then including any items that did not exhibit DIF in an expanded anchor and repeating the process until all remaining items exhibited significant DIF.

The results obtained by Y. Chen et al. (2019) and Y. Chen et al. (2020) varied across scales: For some scales, only one or two items out of nine or 11 exhibited DIF between language / cultural groups; for other scales almost all of the items exhibited DIF when anchored by the single item that best represented the target latent variable. In the former case, removal of a single offending item or

two yielded comparable measurement across languages / cultures, but in the latter case, the results converge to a failure of factorial invariance in the classical factor-analytic sense: It appeared that cultural differences were such that some latent variables simply did not exist in the same form in the two cultures, making comparable measurement impossible.

DIF as Random Effects, Alignment, and Regularization

A special issue of *Frontiers in Psychology* introduced by Van De Schoot et al. (2015, p. 1) extended the use of random effects models from factorial invariance analysis to DIF analysis; Van De Schoot et al. (2015, p. 1) summarized the contents: "Furthermore, our special issue contains two extensions of approximate MI to the field of IRT (see also Fox and Verhagen, 2010). Instead of using substantive prior distributions as in the Bayesian approximate MI method, the method described by Fox establishes a measurement scale across countries and conceptualizes country-specific non-invariance in item parameters as random deviations through country-specific random item effects. In such conceptualization cross-group comparisons can still be made even in the presence of non-invariant items. ... Another contribution to our special issue by Muthén and Asparouhov (2014) concerns the use of the alignment method (see also Asparouhov and Muthén, 2014) in IRT models, a method which is essential when applying approximate MI. This method minimizes a loss function which makes sure that there are a few large non-invariant measurement parameters instead of many smaller non-invariant measurement parameters, an optimal alignment strategy which resembles the rationale underlying rotation of factor solutions in EFA." All of this anticipated the use of regularization in IRT, another way to use data analysis to identify zero or near-zero group differences.

The use of *regularization* in the context of multiple regression was suggested by Tibshirani (1996); specifically, the *least absolute shrinkage and selection operator* (lasso) is intended to provide perform predictor variable selection in multiple regression by shrinking some small regression coefficients to zero. The idea has subsequently been applied much more broadly, including in IRT-based DIF analysis, in which the model is parameterize so that differences between the groups' item parameters are themselves parameters, and then those differences, representing DIF, shrink to zero in the estimation if there is no DIF. Magis, Tuerlinckx, and De Boeck (2015) and Tutz and Schauberger (2015) used the lasso penalty to examine DIF with the one-parameter logistic model, and Bauer, Belzak, and Cole (2020) and Belzak and Bauer (2020) examined its performance with the 2PL model. S. M. Chen, Bauer, Belzak, and Brandt (2022) examined a Bayesian alternative to the lasso, *spike and slab* priors, in the context of DIF detection using the moderated nonlinear factor analysis model.

Using the approach described by Belzak and Bauer (2020) and implemented in Belzak's (2021) software, DIF effects for each item's *a* and *b* parameters were estimated for the spelling test example, shrinking to exactly zero for items that appear to have no DIF. The column labeled $dif_{Regularized}$ in Table 2 contains lasso estimates of *b*-DIF for some of the items; negative values represent DIF favoring males, positive values for items that favor females, and zero when the algorithm indicates no DIF. In addition to the items listed in Table 2 that exhibit DIF according the LR tests with all-other-items as anchor, the lasso regularization analysis estimates non-zero DIF for the words *regress*, *gravitate*, *mariner*, *remember*, *satin*, *archer*, *inspiration*, *pleadingly*, and *zyzygy*. The *b*-parameter locations of those items are shown in Figure 2, which is the same as Figure 1 with different items labeled: The lower seven items in Table 2 are labeled in Figure 2 in black; those items exhibit significant DIF according the to the all-other-items anchored LR procedure, but not by regularization (or the robust method, yet to be described). The additional items labeled in blue

in Figure 2 are flagged for DIF by the regularization procedure, those labeled in red by a robust procedure, and those labeled in magenta by both.

Robust Estimation of the Group Difference

Harking back to the origins of IRT-based DIF analysis, Wang, Liu, and Liu (2022) and Halpin (2022a) have recently proposed methods to use separate IRT calibrations within each of two groups, along with robust estimation of the scaling parameters A and B, to produce modernized versions of Lord's (1977, 1980) analyses that are capable of detecting relatively large proportions of DIF items. Wang et al. (2022) explicitly used plots like those shown in Figures 1 and 2, and similar graphics made using IRT slope (a) parameters, and robust (least trimmed squares, or LTS) estimation of the parameters A and B of the reference line. After experimenting with several combinations using simulated data, Wang et al. (2022, p. 7 of preprint) settled on a system that involved "(1) determining the slope of the reference line of a-DIF based on the a parameters only and (2) fixing the slope of the reference line of b-DIF as the inverse of estimated slope for the reference line of a-DIF and determining the intercept of the reference line of b-DIF using only the b parameters." Wang et al. (2022) worked out the mathematics required to use the vertical deviations of the points from the reference line in graphics like Figures 1 and 2 as DIF test statistics. The found in simulation that this robust method outperformed LR tests using all other items as the anchor when the proportion of DIF item increased up to 50 percent, at which point even the robust system began to break down.

Halpin (2022a) proposed a somewhat different method for robust estimation of the scaling parameters A and B: Instead of explicitly estimating a reference line in a graphic like Figures 1 and 2, he recast the algebra for the IRT differences in intercepts, and subsequently differences in slope parameters, so that selection of the scaling parameters A and B involved robust estimation of location instead of the slope and intercept of a line. Then Halpin used a redescending M-estimator (the Tukey bisquare) to estimate A and B. The bisquare weights computed in this process are a kind of outlier detection: Outliers are weighted zero in the computation of the weighted estimate of location. Outlying differences between groups' item parameters for the same item are DIF, so the weights became an indicator of DIF. Halpin (2022a) also developed test statistics, including a joint Wald test for the 2PL model with two d.f. for the slope and intercept that is comparable to the LR and Wald tests discussed in previous sections.

Halpin's (2022b) method applied to the spelling data yielded the Wald test statistics and weights in the columns labeled $X^2_{Robust}(2)$ anbd Wt_{Robust} in Table 2. The weights are zero (indicating DIF) for all but two of the items in Table 2; the robust test statistics also tend to be greater than the 6.0, the $\alpha = 0.05$ critical value for a χ^2 distribution with 2 d.f., but not all meet the false discovery rate criterion.

As mentioned above, items labeled in Figure 2 in red are flagged for DIF by the robust procedure, and those labeled in magenta by both the regularization and robust algorithms, but neither set of items is flagged for DIF by the LR tests with all-other items as anchor and BH correction for multiplicity. Items with labels in black (in either Figure 1 or 2), exhibiting significant DIF after BH correction of the LR tests with all other items as anchor, tend to be clustered near the center of the scatterplot; that is, their b parameters are near zero, which means they have the smallest standard errors among the bs. Items labeled in color in Figure 2 tend to be farther from the center; those items have b parameters with larger standard errors, making them less likely to be significantly different between groups, for the same degree of deviation from the reference line. The robust and regularized algorithms consider the differences between the b (and a) parameters with respect to

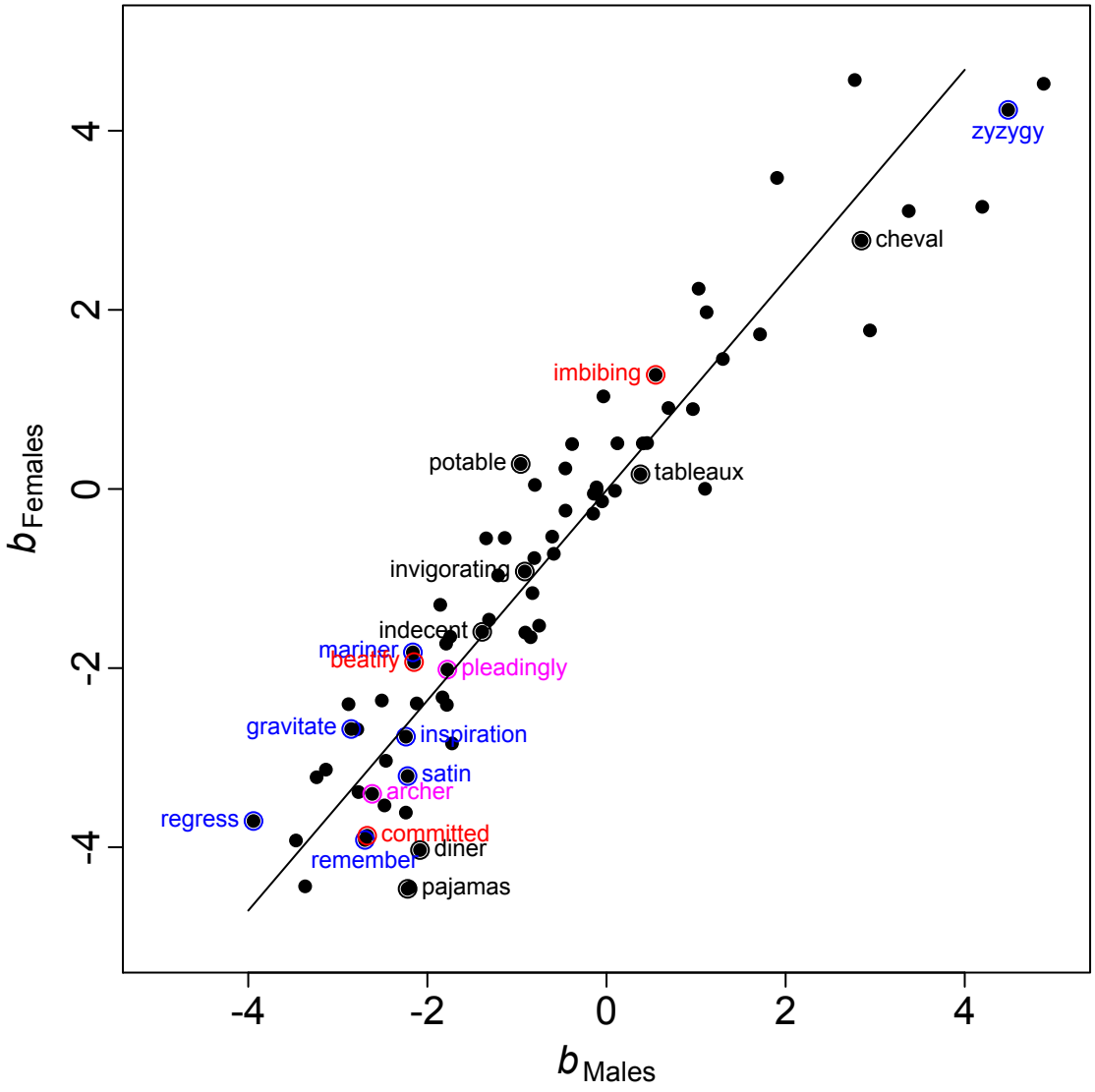

Figure 2

2PL difficulty ($b_{Females}$) parameters of the spelling items for female examinees plotted against difficulty (b_{Males}) parameters of the items for male examinees. Items represented by points above and to the left of the diagonal principal axis reference line are relatively more difficult for female examines to spell, while items with points below and to the right of that line are more difficult for males. The lower seven items in Table 2 are labeled in black. The additional items labeled in red are flagged for DIF by the robust procedure, those labeled in blue by the regularization procedure, and those labeled in magenta by both.

reference lines based on different subsets of the items than the Wald or LR tests; the latter use all (or all other) items. This may account for the fact that the robust and regularized procedures flag some items that do not exhibit DIF when tested for significance by the Wald or LR tests.[20]

A small number of items labeled in Figure 2 are very near the reference line: *inspiration*, *pleadingly*, and *tableaux* draw the eye. The *b* parameters plotted in Figure 2 are the 2PL estimates, and those items have similar *b* parameters associated with somewhat different slopes. As noted above, none of the slope differences between males and females in these data are significant using any of the testing algorithms. When the slopes of those items are constrained to be exactly equal, the *b* parameters for those items shift to become substantially different. The important thing is that the trace lines are different either way.

Discussion: What have we learned?

Factorial Invariance and DIF

A prerequisite for either DIF analysis or the analysis of factorial invariance is *configural* or *pattern* invariance: Do the two groups have the same factor structure? Tests of parameter equality make little sense in the absence of configural invariance. Historically in applications of factor analysis, configural invariance was a real question, and may still be today. However, in a contemporary CFA context, researchers are often fairly confident in the general structure of the data before they are collected. And in the context of DIF analysis, the variables are usually the items on a measurement instrument that is assumed to be essentially unidimensional, so configural invariance is basically the assumed unidimensional model. As IRT-based DIF analysis is done more with multidimensional models, certainly configural invariance will become more of a question, but that is rare now.

Given configural invariance, the contemporary wisdom is that factorial invariance analysis proceeds in steps based on categories of model parameters (Millsap & Meredith, 2007; Putnick & Bornstein, 2016): First comes evaluation of *weak* or *metric* invariance, computing a statistical test for the equality of factor loadings between groups. There can be partial weak (or metric) invariance, if only some of the factor loadings differ between groups. Second, for variables (items) for which the loadings are constrained equal, *strong* or *scalar* invariance is tested by constraining the intercept parameters to be equal, and computing a statistical test of that equality constraint. Partial scalar invariance is a possibility. Finally, *strict* or *residual variance* invariance can be tested with the constraint that the unique variances to be equal for the two groups. This sequence is represented in the lower right hand quadrant of Figure 3.

In contrast, in the DIF literature, the focus is on the items (or variables) first, and the parameters of the model second (Thissen, 2017, p. 86): "The items are divided into an anchor set (items that are assumed to have no DIF), and a studied item (or items). DIF is most often tested one item at a time. When an item is tested for DIF using parametric IRT models, the between-group differences for all of its parameters are tested simultaneously, then perhaps separately (by classes of parameters) within each item." That sequence is represented in the upper left hand quadrant of Figure 3 for the 2PL model[21], or in the upper right hand quadrant if the linear-normal factor model is used for the IRT analysis. While most DIF analyses in the literature are represented schematically

[20] Thanks to Peter Halpin for pointing out this possibility, which deserves future research in some less anecdotal context.

[21] In Figure 3 "2PL" stands in for any IRT model; that reference is explicit for convenience in specifying the parameters *a* and *b*. Other IRT models have different parameters, so a general statement would not be so compact.

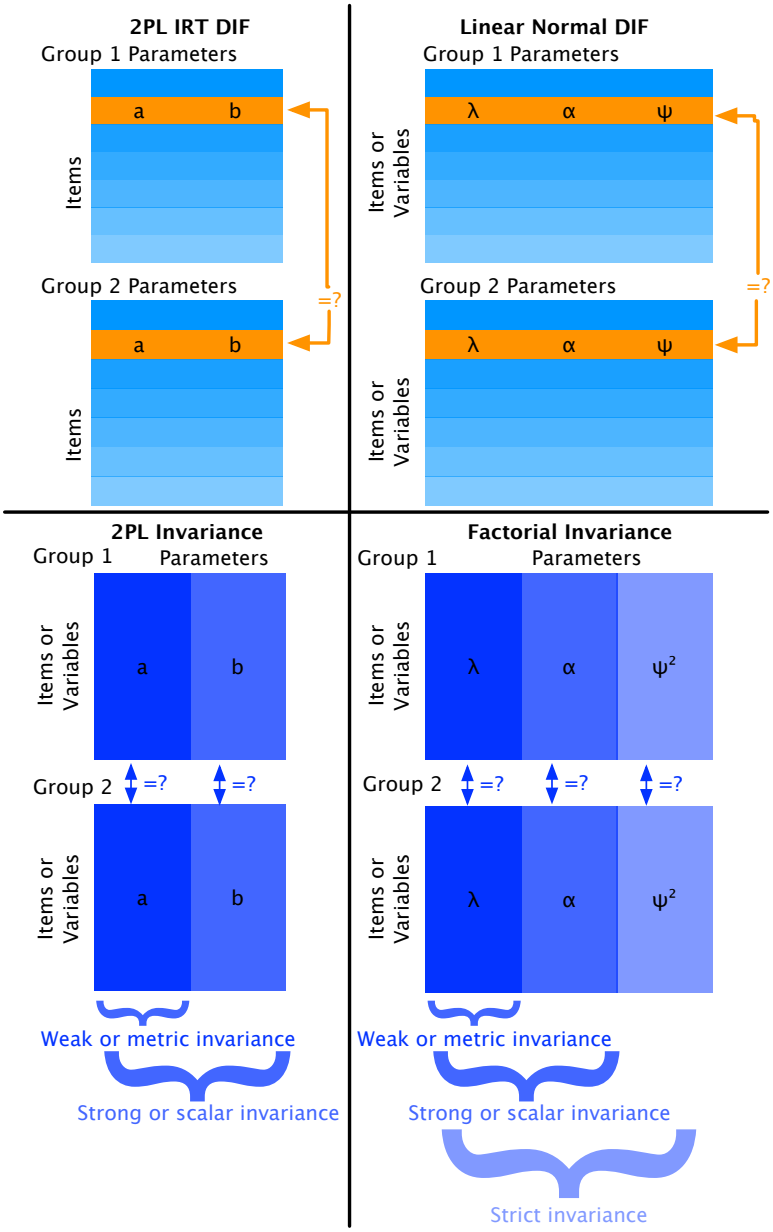

Figure 3

Upper left quadrant: Schematic depiction of IRT-based DIF analysis using the 2PL model, checking the equality of the parameters for a single item (orange) between groups using the other items (cyan) as the anchor. Upper right quadrant: Parallel schematic depiction of CFA-based DIF analysis using the linear-normal factor model, jointly checking equality of each item's parameter set (λ_i, α_i, σ_i^2), repeating item-by-item. Lower left quadrant: Schematic depiction of the use of the IRT 2PL model for the analysis of factorial invariance, testing first equality across groups of the slope parameters as a set, then the thresholds as a set. Lower right quadrant: Schematic depiction of conventional analysis of factorial invariance, checking first the equality of factor loadings between groups, then adding the constraint of equal intercepts, and finally also adding the constraint of equal specific variances.

by the upper left hand quadrant of Figure 3, there have been uses of the linear-normal factor model as an IRT model; examples include studies by Raykov, Marcoulides, and Millsap (2013), Thissen (2017), and Y. Chen et al. (2020). IRT analyses sometimes slip over into the lower left quadrant of Figure 3; examples are studies by Thissen et al. (1986), Sharp et al. (2019), and Y. Chen et al. (2019).

The point of this section is that whether one does statistical testing by blocks of parameters (as in the lower half of Figure 3) or by item (or variables, as in the upper half of Figure 3) is a choice made by the investigator that depends on the context and research question. Traditional invariance analysis is about the veracity of the factor model as a representation of the structure of the data across groups; for continuous data the linear normal factor model may be used, while one might choose some IRT model for categorical responses. DIF on the other hand is all about partial invariance, which does not get much attention in the factorial invariance literature.[22] But if the researcher is confident that a latent variable is measured by a set of items or variables, but wants to identify and set aside any items or variables that might artifactually induce group differences, then the item-by-item tests schematically represented in the upper half of Figure 3 accomplish the goal. The responses to test items may be in categories or on some continuous scale.

The Order of Parameter Tests

In either factorial invariance or DIF analysis, there is universal agreement that the first parameter equality constraints and tests across groups are of the loadings or slope (a) parameters, either as blocks or within items, followed by tests of the intercepts or thresholds conditional on equal slopes, again either as blocks or within items. That is because tests of intercept parameters that are *not* conditional on equal slopes are meaningless, just as the case with traditional analysis of covariance: With unequal slopes, the model lines or curves cross, and so a test of equality is only a test of whether the crossing point (where there is no difference between the groups) happens to be near zero on the x-axis. The third (block of) test(s) in the linear normal model, the test of the unique variances, has no parallel in IRT models as they are usually parameterized.[23]

Ever since orders-of-statistical-tests were suggested by Jöreskog (1971) and Sörbom (1974), in the conventional analysis of factorial invariance the factor (or latent variable) means and covariance matrices have been left free to vary across groups while while tests of loadings, intercepts, and unique variances are done; that is the same as has been done in DIF analysis, leaving the groups means and variances free to reflect *impact*. It is reasonably clear that different factor (or latent variable) means and variances across groups are results, not DIF or a lack of invariance. But the latent variable covariances, or correlations, present a more complicated question.[24]

[22] At the International Meeting of the Psychometric Society at the University of Maryland in July of 2023, Michael Edwards observed that a briefer summary of this is that "there is a different mindset that is reflected in how easy it is to do 'partial invariance.' Item response theorists fully expect to find DIF, so consruct a system that can still function in practice (an important characteristic) even if DIF exists and items cannot be removed. Structural equation modelers test to determine a level of invariance and move on. So, more pithily, IRT folks expect to find DIF, SEM folks hope they won't."

[23] There are parameterizations of some categorical response models that have a within-person-and-item discriminal dispersion parameter separate from the slope, but multi-group data and some cross-group equality constraints on the threshold parameters are required to identify those additional parameters. Those features are not implemented in commonly-available IRT software and such tests are not usually considered in IRT-based DIF analyses.

[24] There is very little literature on multidimensional IRT-based DIF analysis. What has appeared has used orthogonal multidimensional IRT models (Fukuhara & Kamata, 2011; Bulut & Suh, 2017), so the question of what to do with the

Are Latent Variable Correlations Part of Metric Invariance or DIF?

This question is raised when we recall from the decades-old literature on (exploratory) factor analysis, and now from textbooks, that rotation of a factor solution transforms any correlation among the factors into cross-loadings (for orthogonal solutions), and reverses the process to move cross-loadings into inter-factor correlations (for oblique rotations). So if latent variable correlations are interchangeable with loading or slope parameters, should tests of equality of the factor correlations be performed with the loading / slope tests? That is, first in the sequence? And not after tests of the intercepts / thresholds?

Answers to these questions can be developed by thinking about the properties of the scores one may want to compute to characterize the latent variables for individuals. Two cases can be distinguished:

I. Collected Unidimensional Measures. This would be for an *independent clusters* multiple-factors model, with each item associated with one (and only one) factor. Then one could for some purposes compute unidimensional scores one factor at a time. If the factors are correlated, this may be less statistically efficient than using all the data, but the latent variables are defined in this case solely by their relations with a subset of the observed indicators. The latent variables' correlations with each other are consequent to that definition and may vary across groups or situations. This is already what is done with any collection of unidimensional tests collected into a set without joint analysis.

In this case the correlations among the latent variables are consequences or results, and analyses of DIF for individual parameters, or blocked analysis of factorial invariance, would proceed in the order suggested by the factor analytic literature: 1) Loadings / slopes, then 2) intercepts / thresholds, then 3) residual variances if the data are continuous and those are included in the model. All tests would be done with free factor means and variance/covariance matrices for all but a reference, standardizing group. One could add a test or examination of whether the correlational structure of the latent variables differed, but that would be a result after the fact.

II. Explicitly Multidimensional Measures. In this case one anticipates computing or using the scores as a multidimensional set, or a multidimensional score. In this case, if the latent variables are correlated, or if there are cross-loadings, responses to items associated with any of the latent variables may contribute to scores (or measurement) for all of the latent variables. The latent variables are defined in this case both by their relations with the observed indicators and their correlations with each other.

In this case the correlations among the latent variables are integral to the *definition* of the latent variables, as much as the slopes / loadings. So the factor correlations should be grouped with the loadings / slopes in DIF tests (within items), or blocks of parameters for the analysis of factorial invariance. That is, the order would be: 1) Loadings / slopes and factor intercorrelations, then 2) intercepts / thresholds, and then 3) residual variances if the data are continuous and those are included in the model. All tests would be done with free factor means and variances for all but a reference, standardizing group. Note that to do the analyses this way, the latent variable correlations must be separated from the latent variable variances, not treated as a unit, in a variance-covariance matrix. Li, Kleinbort, Thissen-Roe, and Szary (2023) and Kleinbort, Li, Thissen-Roe, and Szary (2023) recently described analyses done in this way.

It is up to the investigator to decide which of the those cases apply.

---

latent variable correlations has not arisen.

The Designated Anchor Redux

A theme of this discussion is that a number of important aspects of the analysis of factorial invariance or DIF are choices or judgments made by the researcher in the service of the context and research questions. That is a general principle of the application of latent variable models. Nowhere is that more salient than in the choice of a set of anchor items for DIF analysis, or, by extension, the set of items or observed variables defining each latent variable in the context of partial factorial invariance.

Recent developments may appear to make anchor selection part of an automated statistical procedure for DIF, for example using alignment or regularization or robust estimation. But it is good to remember Camilli's (1993) observation that "Internal methods of DIF are ipstative" (p. 409). That is, extant purely statistical methods of anchor selection select an average (in some sense) difference between the focal and reference groups as the difference on the latent variable, and deviations from that, positive and negative, become DIF. That can be a bad answer, most obviously in the case in which half the items exhibit one difference between the two groups and the other half the items a different difference, as was the case in the analysis reported by Sharp et al. (2019); the average in that case agrees with none of the observed variables. This is not so much an unusual story as it is a reminder that researchers should always use expert judgment to decide whether the set of items used as the anchor represents measurement of the intended latent variable. External validity evidence could support such judgment. Thissen et al. (1993, p. 102) emphasized the value of a "designated anchor", and that idea is reiterated here.

This idea generalizes to partial factorial invariance as well: Algorithms like alignment or regularization or robust estimation of the reference line match up some central (in some sense) set of variables, leaving other items to exhibit some degree of lack of invariance across groups. However, researchers should remember that it may be that central set represents the intended latent variable less well than some (possibly smaller) block of the items that appear to lack invariance. Analyses leading to partial factorial invariance are exactly the same as DIF analyses, and users of the two types of algorithms can learn one from the other. De Boeck (2023) reminded us in a recent presenation that a majority of items, or even an entire test, may be biased, and DIF analysis using a majority-rules anchor will not show that.

Conclusion

Jöreskog (1971, p. 425) foreshadowed one argument made here, that statistical analysis alone does not provide the answers, when he obtained two different solutions in the very first example of the uses of LR statistical tests for these procedures; he described the two solutions as follows: "One is that the whole factor structure is invariant over populations with a three-factor solution of a fairly complex form. The other is to represent the tests in each population by three factors of a particularly simple form, but these factors have different variance-covariance matrices in the different populations." The three latent variables in that example were defined somewhat (perhaps slightly) differently by the two competing solutions. Jöreskog suggested that more data could settle the issue empirically; that may be the case. However, it is a good reminder that with any given set of data and analyses, ultimately the data analyst must choose.

References

Ahmavaara, Y. (1954). Transformation analysis of factorial data and other new analytical methods of differential psychology with their application to Thurstone's basic studies. *Annales Academiae Scientiarum Fennicae, Sarja-Ser. B*, *88*(2).

Angoff, W., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95-106. doi: 10.1111/j.1745-3984.1973.tb00787.x

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 1-14. doi: 10.1080/10705511.2014.919210

Bauer, D. J., Belzak, W., & Cole, V. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 43-55. doi: 10.1080/10705511.2019.1642754

Belzak, W. (2021). regDIF: Regularized differential item functioning [Computer software manual]. Pittsburgh, PA. Retrieved from `https://github.com/wbelzak/regDIF/` (version 1.0.0)

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*, 673-690. doi: 10.1037/met0000253

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D., Zimowski, M. F., & Thissen, D. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*, 197-211. doi: 10.1111/j.1745-3984.1997.tb00515.x

Bolt, D. M., Hare, R. D., Vitale, J. E., & Neman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist–Revised. *Psychological Assessment*, *16*, 155-168. doi: 10.1037/1040-3590.16.2.155

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*(51), 1-14. doi: 10.3389/feduc.2017.00051

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466. doi: 10.1037/0033-2909.105.3.456

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO version 2: Flexible, multidimensional, multiple categorical IRT modeling [Computer software manual]. Chicago, IL: Scientific Software International.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do test bias procedures obscure test fairness issues? In P. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chalmers, R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29. doi: 10.18637/jss.v048.i06

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors

for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 122-139. doi: 10.1080/10705511.2021.1948335

Chen, Y., Daughters, S., Thissen, D., Salcedo, S., Anand, D., Chen, L., ... Su, L. (2019). Cultural differences in environmental reward across individuals in China, Taiwan, and the United States. *Journal of Psychopathology and Behavioral Assessment*, *41*, 507-523. doi: 10.1007/s10862-019-09743-0

Chen, Y., Thissen, D., Anand, D., Chen, L., Liang, H., & Daughters, S. (2020). Evaluating differential item functioning of the Chinese version of the Behavioral Activation for Depression Scale (C-BADS). *European Journal of Psychological Assessment,*(36), 303-323. doi: 10.1027/1015-5759/a000525

De Boeck, P. (2023, July 25-28). *Pervasive DIF and DIF detection bias.* [Conference presentation]. International Meeting of the Psychometric Society 2023, College Park, MD, United States.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin*, *95*, 134-135. doi: 10.1037/0033-2909.95.1.134

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19-29. doi: 10.1037/0021-9010.72.1.19

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the mini-mental state examination. *Medical Care*, *44*, S134-S142. doi: 10.1097/01.mlr.0000245251.83359.8c

Edwards, M., & Edelen, M. (2009). Special topics in item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (p. 178-198). London: Sage Publications.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*, 23-25. doi: 10.1207/s15327574ijt0501_3

Fox, J.-P., & Verhagen, A. (2010). Random item effects modeling for cross- national survey data,. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (p. 467-488). London: Routledge Academic. doi: 10.4324/9781315537078-19

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*, 604-622. doi: 10.1177/0146621611428447

Guilford, J. P., Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York, NY: McGraw-Hill.

Halpin, P. (2022a, July). *Differential item functioning via robust scaling.* (Unpublished ms.)

Halpin, P. (2022b). R package for differential item functioning using robust statistics [Computer software manual]. Chapel Hill, NC. Retrieved from `https://github.com/peterhalpin/robustDIF`

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), (p. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horn, J., & McArdle, J. (1992). A practical guide to measurement invariance in research on aging. *Experimental Aging Research*, *18*, 117-144. doi: 10.1080/03610739208253916

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417-441. doi: 10.1037/h0071325

Houts, C. R., & Cai, L. (2020). flexMIRT® user's manual version 3.52: Flexible multilevel multidimensional item analysis and test scoring [Computer software manual]. Chapel Hill, NC: Vector Psychometric Group.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426. doi: 10.1007/BF02291366

Jöreskog, K., & Sörbom, D. (1974). *LISREL III [Computer software]*. Chicago, IL: Scientific Software International, Inc.

Jöreskog, K., & Van Thillo, M. (1972, December). *LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables* (ETS Research Bulletin No. RB-72-56). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1972.tb00827.x

Kaiser, H. F. (1958). The varimax criterion for analytical rotation in factor analysis. *Psychometrika*, *23*, 187-200.

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*, 291-312. doi: 10.1207/s15324818ame0804_2

Kleinbort, A., Li, A., Thissen-Roe, A., & Szary, J. (2023, July 25-28). *Global validity of assessments: Location and currency effects.* [Conference presentation]. International Meeting of the Psychometric Society 2023, College Park, MD, United States.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (Second ed.). Springer. doi: 10.1007/978-1-4757-4310-4_6

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (p. 362-412). New York, NY: Wiley. doi: 10.2307/2571672

Li, A., Kleinbort, A., Thissen-Roe, A., & Szary, J. (2023, July 25-28). *Are we playing the same game? Translating fairness content.* [Conference presentation]. International Meeting of the Psychometric Society 2023, College Park, MD, United States.

Lord, F. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (p. 19-29). Amsterdam: Swets and Zeitlinger.

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

Loyd, B., & Hoover, H. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement*, *17*, 179-193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 11-135. doi: 10.3102/1076998614559747

Marco, G. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160. doi: 10.1002/j.2333-8504.1977.tb01136.x

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*, 177-185. doi: 10.1007/BF02289699

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

Millsap, R., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (p. 131-152). Lawrence Erlbaum Associates Publishers. doi: 10.4324/9780203936764

Muthén, B., & Asparouhov, T. (2014). Irt studies of many groups: the alignment method. *Frontiers in Psychology*, *5*, 978. doi: 10.3389/fpsyg.2014.00978

Orlando, M., & Marshall, G. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, *14*, 50-59. doi: 10.1037/1040-3590.14.1.50

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. doi: 10.1016/j.dr.2016.06.004

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502. doi: 10.1007/BF02294403

Raykov, T., Marcoulides, G., & Millsap, R. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, *73*, 713-727. doi: 10.1177/0013164412451978

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552-566. doi: 10.1037/0033-2909.114.3.552

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Reider (Eds.), *Equivalence in measurement: Research in management* (p. 25-50). Greenwich, CT: Information Age.

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2022). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*, 2101-2113. doi: 10.3758/s13428-021-01689-0

Sharp, C., Steinberg, L., Michonski, J., Kalpakci, A., Fowler, C., & Frueh, C. (2019). DSM borderline criterion function across age groups: A cross-sectional mixed-method study. *Assessment*, *26*, 1014-1029. doi: 10.1177/1073191118786587

Sharpe, L. K., & Peterson, R. A. (1971). A comparison of two approaches to the analysis of personality differences. *The Journal of Psychology*, *79*, 257-262.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194. doi: 10.1007/bf02294572

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229-239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Stark, S., Chernyshenko, O., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292-1306. doi: 10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-107. doi: 10.1086/

209528

Steinberg, L., Sharp, C., Fowler, C., & Frueh, C. (2016, October). *When every item exhibits DIF relative to the same anchor: An illustration of the interpretation of DIF*. Presentation at the annual meeting of the Society for Multivariate Experimental Psychology, Richmond, VA.

Steinberg, L., & Thissen, D. (1996, 03). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*, 81-97. doi: 10.1037/1082-989X.1.1.81

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210. doi: 10.1177/014662168300700208

Thissen, D. (2001). IRTLRDIF v2.0b—Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software documentation] [Computer software manual]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D. (2017). Similar DIFs: Differential item functioning and factorial invariance for scales with seven ("plus or minus two") response alternatives. In L. A. der Ark, D. M. Wilberg, S. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative Psychology - The 81st Annual Meeting of the Psychometric Society, Asheville, North Carolina, 2016* (p. 81-91). Springer.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118-128. doi: 10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (p. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thomson, G. H. (1950). *The factorial analysis of human ability* (4th ed.). University of London Press Ltd.

Thomson, G. H., & Ledermann, W. (1939). The influence of multivariate selection on the factorial analysis of ability. *British Journal of Mathematical and Statistical Psychology*, *29*, 288-305. doi: 10.1111/j.2044-8295.1939.tb00919.x

Thurstone, L. L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press. doi: 10.1037/10018-000

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, IL: The University of Chicago Press. doi: 10.1037/h0069792

Thurstone, T. G. (1980). *Chicago & Chapel Hill Recollections.* [Speech audio recording], Chapel Hill, NC: L.L. Thurstone Psychometric Laboratory.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, *58*, 267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*, 21-43. doi: 10.1007/s11336-013-9377-6

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

van der Linden, W. J. (Ed.). (2021). *Handbook of item response theory* (Vol. One: Models). Boca Raton, FL: CRC Press. doi: 10.1201/9781315374512

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01064

Wang, W. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education,*, *72*, 221-261. doi: 10.3200/jexe.72.3.221-261

Wang, W., Liu, Y., & Liu, H. (2022). Testing differential item functioning without predefined anchor items using robust regression. *Journal of Educational and Behavioral Statistics*, *47*, 10.3102/10769986221109208.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (p. 281-324). Washington, DC: American Psychological Association. doi: 10.1037/10222-009

Williams, V. (1993). *An evaluation of item response theory for the investigation of differential item functioning* (Unpublished doctoral dissertation). The University of North Carolina at Chapel Hill, Chapel Hill, NC.

Williams, V. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education*, *10*, 253-267. doi: 10.1207/s15324818ame1003_4

Wood, R., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters [Computer program]* (ETS Resesrch Memorandum No. RM 76-6). Princeton, NJ: Educational Testing Service.

Woods, C. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57. doi: 10.1177/0146621607314044

Woods, C., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group irt. *Educational and Psychological Measurement*, *73*, 532-547.

Yang, F., Yang, K. H. K. M. C., Ocepek-Welikson, K., Kleinman, M., Morales, L. S., Hays, R. D., ... Teresi, J. A. (2011). A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos. *Psychological Test and Assessment Modeling*, *53*, 440-4670.

Zimmer, F., Draxler, C., & Debelak, R. (2022). Power analysis for the Wald, LR, score, and gradient tests in a marginal maximum likelihood framework: Applications in IRT. *Psychometrika*. doi: 10.1007/s11336-022-09883-5