Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 3 | Issue 3 Article 1

2022

Latent Variable Estimation in Factor Analysis and Item Response Theory

David Thissen
University of North Carolina at Chapel Hill

Anne Thissen-Roe
pymetrics

Latent Variable Estimation in Factor Analysis and Item Response Theory

David Thissen [a] and Anne Thissen-Roe [b]

[a]The University of North Carolina at Chapel Hill
[b]pymetrics

Abstract
This essay sketches the historical development of latent variable scoring procedures in the item response theory (IRT) and factor analysis literatures, observing that the most commonly used score estimates in both traditions are fundamentally the same; only methods of calculation differ. Different procedures have been used to derive factor score estimates and latent variable estimates in IRT, and different computational procedures have been the result. Due to differences in the context of score usage, challenges have led to different solutions in the IRT and factor analytic traditions. The needs for bias corrections differ, as do the corrections that have been proposed. While the standard factor analysis model has naturally Gaussian likelihoods, IRT does not, but in IRT normal approximations have been used in various contexts to make the IRT computations more like those of factor analysis. Finally, factor analysis alone has been the home of decades of controversy over *factor score indeterminacy*, while IRT has not, even though the scores in question are the same. That is an artifact of history and the ways the models have been written in the IRT and factor analytic literatures. IRT has never been plagued with questions of indeterminacy, which helps to clarify the position that what is referred to as indeterminacy is not a problem.

Keywords
Factor scores;
Item response theory

It is widely recognized that item response theory (IRT) and factor analysis are members of a family of statistical models that explain the observed covariation among observed variables by appeal to to idea of a smaller number of unobserved, or latent variables (Bartholomew et al., 2011; Skrondal & Rabe-Hesketh, 2004). That being said, IRT and factor analysis have largely separate streams of historical development, because they have been used most often with different data for different purposes: IRT models have been mostly for observed discrete data in two or a few categories; the structural analysis of the variables, or items, has been primarily in the service of test construction to produce individual measurement, which is to say test scores. Factor analysis models originated for observed continuous data; the structural analysis of the variables, that is, the estimation of factor loadings and residual variances, and possibly factor means, has been the primary result in the service of psychological theory. In the factor analytic tradition, more often than not, factor score estimates have been a stepchild, often unwanted.

For the most part neither the IRT nor the factor analysis literature makes the close relationship between factor analysis and IRT clear in any detail. Procedures to compute factor score estimates originated with ideas based on regression (Bartlett, 1937; Thomson, 1935, 1936, 1938; L. L. Thurstone, 1935), and the usual textbook presentation of factor analysis uses those lines of argument. The literature on factor score estimates (that is not preoccupied with factor score indeterminacy [Grice, 2001]) is about the *properties* of the estimates: whether they are conditionally unbiased, whether they have the right variances and correlations with each other or other variables, etc.; for examples see Skrondal and Laake (2001), Devlieger et al. (2015), or Hoshino and Bentler (2013). This presentation

Contact: 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.

is not about those topics, but rather "What if you want factor scores estimates to use like test scores, assuming you know the structural parameters from previous large scale calibration and you probably want to have a method tolerant of observations missing at random?" Mardia et al. (1979), Hoshino and Bentler (2013), Estabrook and Neale (2013), and Loncke et al. (2018) touch on this topic, and Skrondal and Rabe-Hesketh (2004) have a chapter on it, but it has not been salient in the literature on factor analysis.

In this historical essay, we trace the intellectual development of IRT and factor score estimates together in somewhat more detail than in any single source in the existing literature. While we understand that an excellent case can be made for the use of the word *predictions* instead of *estimates* when referring to the single-valued characterizations of the value of the latent variable for individuals (Bartholomew, 1981; Guttman, 1940; Skrondal & Rabe-Hesketh, 2004), and that it is standard usage in the statistical literature to use *prediction* when the inferential target is random, we will use the terms *IRT estimates* and *factor score estimates* for two reasons: (1) It is nearly universal in the IRT literature to call the latent variable values *IRT estimates*, and equally common in the factor analysis literature to discuss *factor score estimates*; changing terminology does little to clarify issues or attenuate controversies. (2) The word *prediction* may carry a connotation for many readers of a future observation, and there is no temporal aspect to the computation of latent variable score estimates.[1] The values of the latent variable are usually thought of as existing in the present, as well as past and future (to some extent). We do understand that, at least from the perspective of frequentist statistics, the values of latent (random) variables are qualitatively different from the (hypothetically) fixed values of structural parameters, for which there are also *estimates*. So the word *estimate* has different meanings in different contexts, even in the same model, but we accept that. Our intention is to provide an explanation that is useful for pedagogy, for both IRT and factor analysis.

We begin by describing the original development of factor score estimates in the 1930s.[2] Then we outline the likelihood principles that were used for the development of IRT scores, and use those principles for factor score estimates to show how they are the same. In the process, we describe ways the factor score estimates can be computed with some of the observations missing at random, again in parallel to standard IRT methods to score around missing item responses. We illustrate the computations with graphics that are rarely used in the factor analytic literature. Finally, we discuss issues that have arisen around latent variable scores, such as their bias, or bias induced in subsequent analyses, the use of normal approximations for IRT computations that render them more similar to the usual factor analytic calculations, and *factor score indeterminacy*.

1 Scores in Linear Factor Analysis: In the Beginning There Was Algebra

1.1 Thomson-Thurstone Regression Estimates

To set context and provide reference notation that will be used intermittently throughout, a modern expression for the multiple factor model for the vector of p observed responses $\mathbf{y}_i$ for person i is

$$\mathbf{y}_i = \mathbf{\Lambda}\mathbf{f}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

in which the $\mathbf{y}_i$ and the vector of k factor scores $\mathbf{f}_i$ are standardized, $\mathbf{\Lambda}$ is $p \times k$ matrix of regression coefficients (or factor loadings) for $\mathbf{y}_i$ on $\mathbf{f}_i$, and $\boldsymbol{\varepsilon}$ is multivariate $N(\mathbf{0}, \mathbf{\Theta})$. $\mathbf{\Theta}$ is the variance-covariance matrix of the residuals (or errors or contributions from *specific factors* or *unique factors*). In classical factor analysis, $\mathbf{\Theta}$ is usually diagonal. We assume throughout that the elements of $\mathbf{\Lambda}$ are sufficiently constrained that the model is uniquely identified; that is, equation 1 represents, in contemporary language, a *confirmatory factor analysis* (CFA) model, with no rotational indeterminacy.

More often this is presented as the consequent model for the covariance matrix among the observations,

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta} \tag{2}$$

in which $\mathbf{\Phi}$ is the covariance (here, correlation) matrix among the factors, for estimation of the parameters in $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ by Wishart maximum likelihood. In traditional factor analysis, $\mathbf{\Phi}$ is usually an identity matrix (that may be omitted entirely from the equation). We are concerned here with estimates of the factor scores $\mathbf{f}$, treating $\mathbf{\Lambda}$, $\mathbf{\Phi}$, and $\mathbf{\Theta}$ as fixed and known.

In the 1930s, however, the model was written differently. Thomson (1936) used the notation

$$\mathbf{z}_i = \mathbf{M}\mathbf{f}_i \tag{3}$$

[1] McDonald (2011) proposed a distinction between *measures* and *predictors* of the value of the latent variables for individuals, both of which we call *estimates* here.

[2] For another description of the early history see the article by Bartholomew et al. (2009).

in which the observations $\mathbf{z}_i$ and the vector of $(k+p)$ factor scores $\mathbf{f}_i$ are standardized, $\mathbf{M}$ is $p \times (k+p)$ matrix of regression coefficients (or factor loadings) for $\mathbf{z}_i$ on $\mathbf{f}_i$. In Thomson's presentation, the *specific factors* are represented explicitly as variables in addition to the common factor scores, and the matrix $\mathbf{M}$ is $[\mathbf{\Lambda}|\mathbf{\Theta}^{\frac{1}{2}}]$, in which $\mathbf{\Theta}^{\frac{1}{2}}$ refers to a diagonal matrix with diagonal elements equal to the square roots of those of $\mathbf{\Theta}$, in the notation of equation 2.

Thomson's representation of equation 2 was

$$\mathbf{R} = \mathbf{M}\mathbf{M}' \tag{4}$$

Thomson's description of his original development of factor score estimates was not motivated by statistical considerations as it might be now, but rather as an algebraic exercise: He assumed the values in $\mathbf{M}$ were known or had been estimated; then premultiplying the left side of equation 3 with an identity matrix in the form of $\mathbf{M}\mathbf{M}'\mathbf{R}^{-1}$, Thomson obtained

$$\mathbf{M}\mathbf{M}'\mathbf{R}^{-1}\mathbf{z}_i = \mathbf{M}\mathbf{f}_i \tag{5}$$

then "dropping the premultipler $\mathbf{M}$ on both sides (although it is not square) we have

$$\mathbf{M}'\mathbf{R}^{-1}\mathbf{z}_i = (\text{estimated})\mathbf{f}_i \tag{6}$$

a matrix equation which can be shown to give the best loadings of the tests $\mathbf{z}$ to measure the factors $\mathbf{f}$ " Thomson (1936, p. 41). The meaning of "best" in this context is only alluded to later in Thomson's 1936 article and clarified in subsequent articles and letters.

From a contemporary perspective, it may be viewed as a curiosity that Thomson's equations could be used to compute estimates for the specific factors as well as the general or group factors. Thomson follows Spearman (1927) in this respect. Modern usage usually only considers estimates of the general or group factors. In a subsequent section we will see how consideration of values for specific factors as well as the general or group factors contributed to decades of controversy over *factor score indeterminacy*.

In modern notation we would write Thomson's equation 6 as

$$\bar{\mathbf{f}}_i = \mathbf{\Lambda}'\mathbf{\Sigma}^{-1}\mathbf{y}_i = (\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda} + \mathbf{\Phi}^{-1})^{-1}\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{y}_i \tag{7}$$

in which the first two terms are direct translations of equation 6, and the mysteriously different final term has been constructed using what is now referred to as the Woodbury identity ("Woodbury matrix identity", 2021) to replace the apparently required inversion of a $p \times p$ matrix with inversion of a (usually much) smaller $k \times k$ matrix (Guttman, 1940; Ledermann, 1939). [Note that the $p \times p$ matrix $\mathbf{\Theta}^{-1}$ is only trivially and notationally a matrix inverse, because $\mathbf{\Theta}$ is diagonal.]

The estimates $\bar{\mathbf{f}}_i$ in equations 6 and 7 are usually referred to as "Thomson-Thurstone regression estimates", because they are the same as L. L. Thurstone (1935, pp. 226-228) proposed in the final chapter of *The Vectors of Mind*.[3] Thurstone only considered estimates of the general or common factors, and explicitly derived the equations as the regression of the factors on the observed tests. At this point in history, no one was using the terms *factor scores* or *factor score estimates*; Thomson wrote about estimating the *factors* and Thurstone simply wrote about x, his notation for what are now called factor scores.[4] In any event, the form of the Thomson-Thurstone regression estimates is easily recognized as standard linear regression of $\mathbf{f}$ on $\mathbf{y}$: the correlations between $\mathbf{f}$ and $\mathbf{y}$ multiplied by the inverse of the covariance matrix among the predictors $\mathbf{y}$. L. L. Thurstone (1935, p. 226) derived the result by explicitly minimizing the summed squared residuals predicting the factor scores from the observed scores.

While there are no cross-references of either Thurstone or Thomson to the other in their 1935-1936 publications (indeed, there are no references at all in Thurstone's six-page chapter on score estimation), it is highly likely that each knew how the other was thinking about factor score estimates. 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). Thomson and Thurstone corresponded and visited to the extent that was possible trans-Atlantic in the 1930s.

1.2 Bartlett Estimates

Bartlett (1937) briefly summarized Thomson's (1936) development of the regression estimates, and then took a step toward the contemporary description of the factor model by separating the common and specific parts. Bartlett (1937, p. 99) wrote "If, however, we adopt the principle

[3] In the interest of cross-literature comparisons to be made later, we use a bar over the estimated quantity to refer to regression or conditional expectation or expected *a posteriori* (EAP) estimates, and a hat to indicate estimates which are (ultimately) maximum likelihood (ML), even in notation that otherwise follows original sources that do not use that convention or those derivations.

[4] The earliest use of the term *factor scores* we have encountered is in the article by Guttman (1940).

... that specific factors should only be introduced only to explain discrepancies between observed scores and postulated general or group factors, we write

$$\mathbf{T} = \mathbf{M}_0\mathbf{F}_0 + \mathbf{M}_1\mathbf{F}_1 \tag{8}$$

where $\mathbf{F}_0$ is the set of general and group factors and $\mathbf{F}_1$ is the specifics. $\mathbf{M}_1$ is thus a diagonal matrix." $\mathbf{T}$ is Bartlett's notation for "a column vector denoting a series of standardized tests", which is to say the observed variables.

Bartlett (1937, p.100) then set up equations to find the minimum of the sum of squares of the specific factors, obtaining "In matrix notation for any number of factors

$$\mathbf{f}_0 = (\mathbf{M}_0'\mathbf{M}_1^{-2}\mathbf{M}_0)^{-1}\mathbf{M}_0'\mathbf{M}_1^{-2}\mathbf{T} \tag{9}$$

This equation represents our estimates for *all* our persons, if more generally we regard $\mathbf{T}$ as the matrix set of all the persons' tests scores."

In contemporary notation we would write

$$\hat{\mathbf{f}}_i = (\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{y}_i \tag{10}$$

which is the usual expression for what are now called *Bartlett factor score estimates*.

In an exchange of letters published in *Nature*, Thomson (1938) and Bartlett (1938) clarified the similarities and differences between their factor score estimate proposals. Thomson (1938, p. 141) wrote that "My formulae were arrived at by the ordinary regression method. Bartlett's estimates and the regression estimates attain different ends, and it is agreed that each method is correct in the right place. The regression estimates minimize the squares of the discrepancies between the estimates and the true values, summed over the population. Bartlett's estimates minimize the squares of a man's *specific* factors, summed over the tests." Citing the matrix algebra in Thomson's letter, Bartlett (1938, p. 609) provided the matrix rescalings that transform regression estimates into his estimates and vice versa:

$$\hat{\mathbf{f}}_0 = \mathbf{K}\bar{\mathbf{f}}_0 \quad , \quad \bar{\mathbf{f}}_0 = \mathbf{K}^{-1}\hat{\mathbf{f}}_0 \tag{11}$$

in which $\mathbf{K} \equiv \mathbf{M}_0'\mathbf{R}^{-1}\mathbf{M}_0$ and $\mathbf{R} \equiv \mathbf{M}_0\mathbf{M}_0' + \mathbf{M}_1{}^2$. Note that for a one-factor model, $\mathbf{K}$ is actually a scalar, the squared multiple correlation coefficient for the factor score predicted by the several indicators. So the left expression in equation 11 is essentially Kelley's (1927, p. 177) regression estimate of the true score, treating the Bartlett estimates as the observed scores. For multi-factor models , $\mathbf{K}$ may be a more complex rescaling of the Bartlett estimates into the regression estimates or vice-versa. Nevertheless, as Thomson (1938, p. 609) pointed out, "*as a vector statistic* one estimate is equivalent to the other." In the case of multiple factors, the diagonal elements of $\mathbf{K}$ are the squared multiple correlations of each successive factor with the indicators as predictors; McDonald (1974) indicates that is one possible measure of factor determination or determinacy.

2 IRT and Factor Analysis: Unidimensional Likelihood-Based Score Estimates

2.1 IRT Scoring

Lawley (1940) published an article on *the estimation of factor loadings by the method of maximum likelihood* (ML). That has little directly to do with the story of factor score estimates, and the use of ML to estimate the structural parameters of the factor analysis model did not come into widespread use until modifications of Lawley's method were implemented in computer applications decades later. However, Lawley (1943)[5] next turned his attention to *problems connected with item selection and test construction*, describing an algorithm to estimate the parameters of what was later to be called the *normal ogive* item response model. Lawley (1943, p. 273) wrote that "we assume that all the items composing a given test are measuring the same ability x ... [T]he probability of a person passing a given item will depend upon his ability as measured on this scale. Ferguson (1942) has made the hypothesis ... that the probability P is given by the relation

$$P = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\alpha}{\sigma}} \exp\left[\frac{u^2}{2}\right] du \tag{12}$$

where x measures the ability of the person and α, σ are constants for a particular item."

α and σ are now referred to as the item parameters in this *normal ogive* IRT model. Lawley (1943, p. 274) proposed dividing the sample into groups (by some unspecified procedure) "such that the variation in ability within each group is sufficiently small for us to be able to neglect it." He further specified that the ability x for each group is known. Then he provided the equations to be solved to do ML probit analysis to obtain estimates of α and σ. There being no practical way to divide the data into the homogeneous groups required, or to know their values of

[5] Lawley's paper was communicated to the Royal Society by Godfrey Thomson.

ability x, Lawley's proposal was not useful, but it planted the seeds of approaching IRT, and computation of its scores, as an estimation problem that could be solved using the method of ML.

Neither Lawley nor the early writers about factor analysis distinguished between fixed unknown values like the item parameters in IRT or the loadings in factor analysis and the latent variables. They simply referred to the underlying construct as the *factor* or x or *ability*. That changed with Paul Lazarsfeld's (1950) chapters in *The American Soldier* series. Lazarsfeld's model was based on functions (or, graphically, lines or curves) describing the probability of a response to an item as a function of the latent (unobserved) variable x. Lazarsfeld (1950b, p. 367) wrote that "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." A "pure test," in Lazarsfeld's language, is a test in which the item responses fit a model with the properties of unidimensionality and local independence. A contemporary statistician sees the same definition of the *factor* in unidimensional factor analysis: a variable that accounts for the interrelationships among the observed variables, or (equivalently) that renders the responses independent conditional on the value of that unobserved variable.

Lazarsfeld (1950b, p. 389) continued "The total sample is therefore characterized by a distribution function $\phi(x)$ which gives for each small interval dx the number of people $\phi(x)dx$ whose score lies in this interval. We can now tell what proportion of respondents in the whole sample will give a positive reply to item i with trace line $f_i(x)$ ..." That is, Lazarsfeld not only made clear that the theory was that there was a latent (unobserved) variable underlying the observed item responses, but also that there were two distinct functions: The population distribution of that latent variable he called $\phi(x)$ and the "trace line $f_i(x)$" for item i. Lazarsfeld described in equations how the joint probabilities of combinations of item responses are modeled as products of the trace lines.

Lazarsfeld (1950a, pp. 460-465) introduced "trace line scores" of response patterns, in what may have been the first description of the ideas of what are now called ML and expected *a posteriori* (EAP) scores. Lazarsfeld (1950a, p. 464) used the terms "*maximum probability score* (MPS)" for the ML estimate and "*expected value score* (EVS)" for the EAP value. Lazarsfeld used linear trace lines, presumably unaware of the then-small psychometric literature on the normal ogive IRT model, and possibly because hand computation was much easier with the equation of a straight line than with an integral expression. But his chapters contained the entire conceptualization of IRT item analysis and scoring.

Lazarsfeld's writing had little visible effect on quantitative psychology at the time, except as it was mirrored by Frederic Lord's (1952) description of ability as an unobserved variable defined by its relationship with item responses.[6] The major point of Lord's monograph was to distinguish between the properties of the unobserved ability variable and observed test scores. Lord (1952, p. 1) wrote:

> A mental trait of an examinee is commonly measured in terms of a test score that is a function of the examinee's responses to a group of test items. For convenience we shall speak here of the "ability" measured by the test, although our conclusions will apply to many tests that measure mental traits other than those properly spoken of as "abilities." The ability itself is not a directly observable variable; hence its magnitude ... can only be inferred from the examinee's responses to the test items.

Lord (1953) expanded on Lawley's ML approach to IRT parameter estimation by proposing that ML could be used to simultaneously provide estimates of the item parameters (as Lawley had suggested) and the values of ability, the latent variable. That idea did not distinguish between the statistical properties of the unobserved person ability variable and the item parameters, even though Lord wrote about them as distinct conceptually. This differed from Lazarsfeld's (1950a, 1950b) chapters that distinguished between estimation of the trace line parameters as the analysis of the structural model, and subsequent computation of trace line scores.

By the 1970s mainframe computers permitted implementation of IRT methods in software, which almost universally followed the suggestion by Birnbaum

[6]Lord was writing his dissertation (Lord, 1952) in New York City at the same time as Lazarsfeld wrote his part of *The American Soldier* as a professor at Columbia. However, it is not clear how much direct influence Lazarsfeld's work might have had on Lord. Lord (1952) did not cite Lazarsfeld; Lord (1953) mentions Lazarsfeld (1950) only in passing.

(1968) that the logistic function

$$T(u_j = 1|\theta) = \frac{1}{1+\exp[-a_j(\theta - b_j)]} \tag{13}$$

be used in place of the normal ogive model for the trace line. In equation 13, $u_j = 1$ is a correct or positive response to item j, a_j and b_j are item parameters, and θ is the latent variable. In the rest of this and subsequent sections, we will use the notation $T(u_{ij}|\theta)$ for *trace line* with a nod to Lazarsfeld.[7]

As a bridge with the earlier normal ogive tradition, Birnbaum pointed out that Haley (1952) had shown that a multiplier of 1.7 applied to the parameter a in equation 13 renders P in equation 12 and T in equation 13 nearly identical with the same parameters; some software embedded 1.7 in the logistic. Equation 13 is written for dichotomous responses, but there are also trace line models for items with polytomous responses with more than two values for u_j.

Thissen and Orlando (2001) and Thissen, Nelson, Rosa, and McLeod (2001) summarize contemporary methods of IRT test scoring. The basic element of the model is the item response function $T(u_{ij}|\theta)$. Most software uses the logistic in equation 13 for dichotomous items, but the normal ogive may also be used; for our purpose here the distinction is not important.

Early computer applications of IRT in the 1970s used ML estimation of the latent variable as the standard method. Under IRT's defining assumption of conditional or local independence, the likelihood of the set of responses $\mathbf{u}_i$ from person i is

$$L(\mathbf{u}_i|\theta) = \prod_j T(u_{ij}|\theta) \tag{14}$$

The mode of equation 14, is the *maximum likelihood* (ML) estimate, as suggested by Lord (1953). That value has been used as an IRT test score, because it can be described as the most likely value of θ given the response pattern $\mathbf{u}$. There is no closed-form solution for the ML estimate in IRT, so the mode is located iteratively, usually with the Newton-Raphson algorithm.

Lazarsfeld (1950b) had pointed out that there must be some density $\phi(\theta)$ for the latent variable itself. Following Lord's (1952, 1953) lead in omitting from the model the idea of the population distribution of the latent variable had some disadvantages in IRT, most notably that all-correct or all-incorrect item response patterns did not have finite ML score estimates. Including the population distribution, a more complete likelihood is

$$L_p(\theta|\mathbf{u}_i) \propto \prod_j T(u_{ij}|\theta)\phi(\theta) \tag{15}$$

Equation 15 is often referred to as the *posterior* density for θ because of the formal analogy of the equation with likelihood-times-prior in Bayesian analysis. That makes the population distribution for the latent variable, $\phi(\theta)$, a *prior* distribution, which it really is not. $\phi(\theta)$ is part of the model for the categorical responses. However, the nomenclature has become so solidified that IRT estimates based on equation 15 are referred to as *a posteriori*.

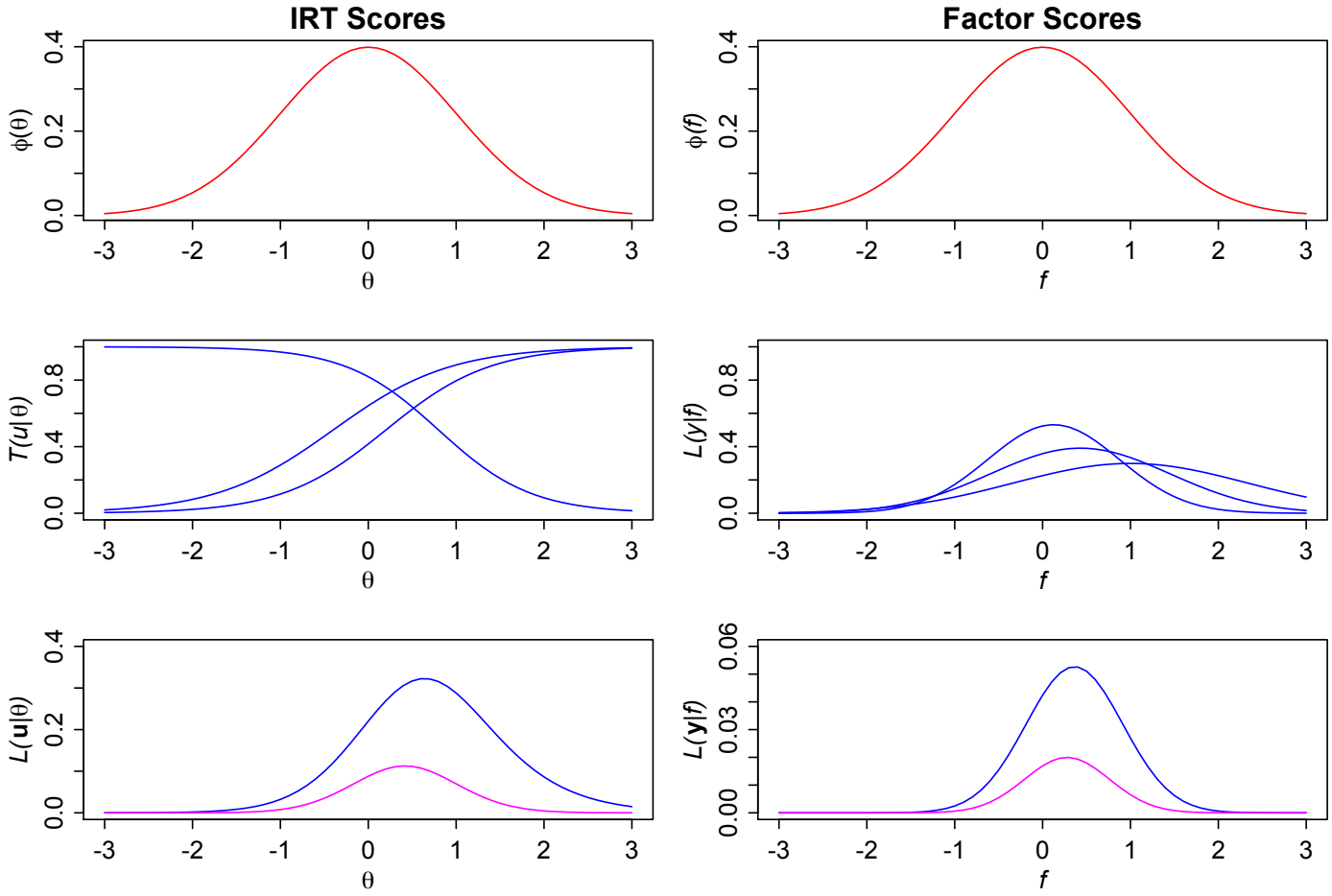
Bock and Mislevy (1982) proposed the use of the expected *a posteriori* (EAP) estimate as an alternative to ML. That idea revisited Lazarsfeld's (1950) EVS, the mean of equation 15, computed by numerical integration. The mode of equation 15, known as the maximum *a posteriori* (MAP) estimate, may also be used. EAP estimates are more computationally intensive than ML or MAP estimates, because the latter can usually be obtained with five or fewer iterative evaluations of the likelihood while EAPs are often computed using 20-50 quadrature points for numerical integration, meaning that many likelihood evaluations. The EAP estimate has the advantage of minimizing squared error, but can be challenging to compute for models with higher-dimensional θ, for which the MAP estimate remains valuable.

From the 1980s onward, most IRT software would follow Lazarsfeld's model, distinguishing between ML estimation of the item parameters, with subsequent computations of score estimates. Using Lazarsfeld's ideas in this way made IRT parallel with factor analysis, which estimates the structural parameters to explain the observed variables' covariation, with factor score estimates computed later if at all.

Unidimensional IRT score computation is illustrated on the left side of Figure 1. The mode of the blue likelihood in the lower panel is the ML estimate. The mode of the magenta "posterior" is the MAP estimate, and the mean of that curve is the EAP estimate. The two modal estimates are computed with an optimization method, usually Newton-Raphson; the EAP estimate is computed using numerical integration.

[7] Use of the notation θ for the latent variable as has become traditional in IRT. This is re-use of the same symbol with two different meanings here, one in the factor analysis model where it is residual variance and the second in IRT where it is the latent variable. We trust context makes the interpretation clear for each appearance.

Figure 1



Left side: IRT scores. Upper panel: The population distribution $\phi(\theta)$, usually $N(0,1)$ for IRT models. Center panel: Three logistic trace lines $T(u|\theta)$, two for correct or positive responses to dichotomous items and one incorrect or negative; $\mathbf{a} = [1.5, 1.7, 1.9]$, $\mathbf{b} = [-0.4, 0.2, 0.8]$. Lower panel: The blue curve is the likelihood $L(\mathbf{u}_i|\theta)$ from equation 14, the product of the three trace lines in the center panel; the magenta curve is the posterior, $L_p(\theta|\mathbf{u}_i)$ from equation 15, which is the blue likelihood times the red population distribution in the upper panel.
Right side: Factor scores. Upper panel: The population distribution $\phi(f)$, often $N(0,1)$ for CFA models. Center panel: Three Gaussian likelihoods $L(y_{ij}|f)$ for three values of $\mathbf{y} = [0.1, 0.3, 0.6]$ and $\boldsymbol{\lambda} = [0.8, 0.7, 0.6]$. Lower panel: The blue curve is the likelihood $L(\mathbf{y}_i|f)$ from equation 18, the product of the three likelihoods in the center panel; the magenta curve is the posterior, $L_p(f|\mathbf{y}_i)$ from equation 19, which is the blue likelihood times the red population distribution in the upper panel.

2.2 Revisiting Factor Analysis

2.2.1 The Likelihood for the One-Factor Model

To parallel unidimensional IRT, we begin with a one-factor version of equation 1 for continuous response y_{ij} for person i and observed variable j,

$$y_{ij} = \lambda_j f_i + \varepsilon_{ij} \tag{16}$$

in which the observations y_{ij} and the factor scores f_i are both assumed to be standardized (hence the absence of an intercept in equation 16). λ_j is the regression parameter (or factor loading) for y_j on f (and also the correlation for standardized y and f), and ε_{ij} is $N(0,\theta_j)$ in which θ_j is the unique, or error, variance for observed variable j. Note that with everything standardized, $\theta_j = 1-\lambda_j^2$ (and the notation θ no longer refers to the latent variable as it did in the preceding section on IRT).

The context for computing factor score estimates is one in which the values of λ_j and θ_j are taken to be fixed and known, just as as the item parameters are for IRT scoring; in both cases those structural parameters are usually previously-estimated with data from large samples.

The (Gaussian) likelihood for response y_{ij}, analogous to the IRT trace line, is

$$\begin{aligned} L(y_{ij}|f_i) &= \phi(y_{ij}|f_i) \\ &= \frac{1}{\sqrt{2\pi\theta_j}} \exp\left[\frac{-(y_{ij}-\lambda_j f_i)^2}{2\theta_j}\right] \end{aligned} \tag{17}$$

Assuming local independence, the (also Gaussian) likelihood of the vector of responses $\mathbf{y}_i$ for person i is

$$L(\mathbf{y}_i|f_i) = \prod_j L(y_{ij}|f_i) \tag{18}$$

Equation 18 is directly analogous to equation 14 of IRT. It ignores the population distribution for simplicity. This is harmless; unlike the situation with IRT, for all patterns of observed responses, finite estimates can still be computed by ML methods. However, factor analysis often makes use of the assumption that the factor scores f are normally distributed in the population. The likelihood then includes the population distribution:

$$L_p(f_i|\mathbf{y}_i) \propto \prod_j L(y_{ij}|f_i)\phi(f) \tag{19}$$

analogous to equation 15 for IRT.

Because both equations 18 and 19 are Gaussian likelihoods, the modes and means are the same, and derivational and computational approaches to compute either provide the same factor score estimates.

2.2.2 Modal (or Maximum Likelihood or ML, or MAP) Estimation

Mardia et al. (1979, p. 274) point out that likelihood-based factor score estimates (ML or MAP or EAP, the latter two of which are identical) are the same as Bartlett's (for ML) and Thomson's (for EAP/MAP) regression-based factor score estimates, and Skrondal and Rabe-Hesketh (2004, p. 239) make the same observation about Bartlett's estimates. Hoshino and Bentler (2013, p. 47) observe about "(1) Bartlett's method and (2) the regression method. The former can be considered the ML estimator of the factor score vector in the fixed effect factor analysis, while the latter can be regarded as the Bayes posterior mean estimator." Estabrook and Neale (2013) discuss the performance of likelihood-based estimates of factor scores, as compared with Bartlett's method. None of those sources provide much detail about their derivation or computation. Thissen and Thissen-Roe (2020) provided some of the detail that follows:

The maximum likelihood, or ML, factor score estimate can be computed by locating the modal value of the likelihood $L(\mathbf{y}_i|f)$ in equation 18, which is the same as the modal value of the log likelihood:

$$\begin{aligned} \ell = \log L(\mathbf{y}_i|f_i) &= \sum_j \log L(y_{ij}|f_i) \\ &= k + \sum_j \frac{-(y_{ij}-\lambda_j f_i)^2}{2\theta_j} \end{aligned} \tag{20}$$

in which k is the log of the norming constant; the maximum value of f is found where the first derivative of ℓ equals zero:

$$\frac{\partial \ell}{\partial f} = \sum_j \frac{\lambda_j(y_{ij}-\lambda_j f_i)}{\theta_j} = 0 \tag{21}$$

Standard ML theory provides the result that the error variance of the estimate is the negative inverse of the second derivative:

$$\frac{\partial^2 \ell}{\partial f^2} = \sum_j \frac{-\lambda_j^2}{\theta_j} \tag{22}$$

For the special case of a single observed variable, Thissen and Thissen-Roe (2020) point out that equation 21 can be solved by visual inspection to yield the mean value of f for an item given the response, $\mu_{f|y_{ij}} = y_{ij}/\lambda_j$; the associated

variance comes from equation 22: $\sigma^2_{f|y_{ij}} = \theta_j/\lambda_j^2$. These can be used to plot graphics for factor score estimation parallel to the left side of Figure 1, as shown on the right side.

The likelihood in equation 18 is a product of Gaussian likelihoods. The mean of a product of Gaussian likelihoods is the average of the means of the component normal distributions, each weighted by the inverse of the associated variance. So the closed form computation of the ML estimate $\hat{f}_i$ is

$$\hat{f}_i = \frac{\sum_j \frac{\mu_{f|y_{ij}}}{\sigma^2_{f|y_{ij}}}}{\sum_j \frac{1}{\sigma^2_{f|y_{ij}}}} = \frac{\sum_j \frac{y_{ij}/\lambda_j}{\theta_j/\lambda_j^2}}{\sum_j \frac{1}{\theta_j/\lambda_j^2}} = \frac{1}{\sum_j \frac{\lambda_j^2}{\theta_j}} \sum_j \frac{\lambda_j y_{ij}}{\theta_j} \tag{23}$$

and the associated error variance is

$$\sigma^2_{\hat{f}_i} = \frac{1}{\sum_j \frac{1}{\sigma^2_{f|y_{ij}}}} = \frac{1}{\sum_j \frac{1}{\theta_j/\lambda_j^2}} = \frac{1}{\sum_j \frac{\lambda_j^2}{\theta_j}} \tag{24}$$

The summations run over all non-missing responses if there are missing data. So if missing responses can be considered missing at random, factor score estimates can be computed skipping over the missing responses just as is the case for IRT θ estimates. We note in agreement with Mardia et al. (1979), Skrondal and Rabe-Hesketh (2004), and Hoshino and Bentler (2013) that equation 23 is the one-factor specialization of equation 10 for Bartlett factor score estimates.

The log likelihood and its derivatives for L_p, which includes a normal population distribution, add terms to equations 20, 21, and 22. A closed form solution to compute analogs to IRT's MAP or EAP is straightforward:

If a $N(0,1)$ population distribution for f is included in the model, that is another Gaussian component in the weighted sum with a mean of 0 and an inverse variance weight of 1, so equation 23 becomes the MAP estimate

$$\begin{aligned} \bar{f}_i &= \frac{\sum_j \frac{\mu_{f|y_{ij}}}{\sigma^2_{f|y_{ij}}} + \frac{0}{1}}{\sum_j \frac{1}{\sigma^2_{f|y_{ij}}} + \frac{1}{1}} \\ &= \frac{\sum_j \frac{y_{ij}/\lambda_j}{\theta_j/\lambda_j^2}}{\sum_j \frac{1}{\theta_j/\lambda_j^2} + 1} \\ &= \frac{1}{1+\sum_j \frac{\lambda_j^2}{\theta_j}} \sum_j \frac{\lambda_j y_{ij}}{\theta_j} \end{aligned} \tag{25}$$

the final expression in equation 25 is given by Thomson (1935, p. 249) as " "estimated" g ... by the Spearman rules or by a regression equation." Spearman (1927, p. xix) provides the weights but not the norming constant. The error variance for the MAP estimate is

$$\sigma^2_{\bar{f}_i} = \frac{1}{\sum_j \frac{1}{\sigma^2_{f|y_{ij}}} + \frac{1}{1}} = \frac{1}{\sum_j \frac{1}{\theta_j/\lambda_j^2} + 1} = \frac{1}{1+\sum_j \frac{\lambda_j^2}{\theta_j}} \tag{26}$$

3 Multiple Factor Analysis Redux: Likelihood-Based Score Estimates

3.1 The Likelihood for the Multiple Factor Model

If the multiple factor model is expressed as equation 1, the likelihood for the data as a function of the factor scores (assuming $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ are known) is

$$L(\mathbf{y}_i|\mathbf{f}_i) = \frac{|\mathbf{\Theta}|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i)'\mathbf{\Theta}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i)} \tag{27}$$

The log likelihood is then

$$\ell = \log L(\mathbf{y}_i|\mathbf{f}_i) = k - \tfrac{1}{2}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i)'\mathbf{\Theta}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i) \tag{28}$$

If a standard normal population distribution for $\mathbf{f}$ is included in the model, with correlation matrix $\mathbf{\Phi}$ among the factors, Bayes' theorem can be used to provide the likelihood of the factor scores given the data:[8]

$$\begin{aligned} L_p(\mathbf{f}_i|\mathbf{y}_i) \propto & \left[\frac{|\mathbf{\Theta}|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i)'\mathbf{\Theta}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i)} \right] \\ & \left[\frac{|\mathbf{\Phi}|^{-\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} \mathrm{e}^{-\frac{1}{2}\mathbf{f}_i'\mathbf{\Phi}^{-1}\mathbf{f}_i} \right] \end{aligned} \tag{29}$$

the log likelihood is then

$$\begin{aligned} \ell_p &= \log L_p(\mathbf{f}_i|\mathbf{y}_i) \\ &= k - \tfrac{1}{2}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i)'\mathbf{\Theta}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\boldsymbol{f}_i) - \tfrac{1}{2}\mathbf{f}_i'\mathbf{\Phi}^{-1}\mathbf{f}_i \end{aligned} \tag{30}$$

[8] The fact that Bayes' theorem is used to reverse the arguments of the likelihood does not necessarily make this *Bayesian* in the senses in which that term is usually used. The population distribution here does not necessarily represent degrees of belief; it has a frequentist interpretation as part of the statistical model for the observed data, just as it does in IRT. This is all *compatible* with Bayesian analysis but it is not, in and of itself, Bayesian.

3.2 Modal Estimation for the Multiple Factor Model

The matrix form of the log likelihood is equation 28; the vector derivative with respect to $\mathbf{f}$ is

$$\frac{\partial \ell}{\partial \mathbf{f}} = \mathbf{\Lambda}'\mathbf{\Theta}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i) \tag{31}$$

and the second derivative is

$$\frac{\partial^2 \ell}{\partial \mathbf{f}} = -\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda} \tag{32}$$

The first derivatives in equation 31 are zero when

$$\hat{\mathbf{f}}_i = (\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{y}_i \tag{33}$$

which was equation 10, the *Bartlett factor score estimates*.

MAP estimates are obtained by adding the derivative components with respect to $\mathbf{f}$ from the standard normal population distribution to equations 31 and 32, as follows:

$$\frac{\partial \ell_p}{\partial \mathbf{f}} = \mathbf{\Lambda}'\mathbf{\Theta}^{-1}(\mathbf{y}_i - \mathbf{\Lambda}\mathbf{f}_i) - \mathbf{\Phi}^{-1}\mathbf{f}_i \tag{34}$$

and the second derivative is

$$\frac{\partial^2 \ell_p}{\partial \mathbf{f}} = -\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda} - \mathbf{\Phi}^{-1} \tag{35}$$

The first derivatives in equation 34 are zero when

$$\bar{\mathbf{f}}_i = (\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda} + \mathbf{\Phi}^{-1})^{-1}\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{y}_i \tag{36}$$

which is the Thomson-Thurstone regression estimate of equation 7. Curiously solving the likelihood's normal equations has brought us to the notational representation of those estimates *after* application of the Woodbury identity by Lederman (1939). There are two easy ways to see that this MAP estimate is also the EAP estimate: The first is to remember that these likelihoods are Gaussian, so the mode and the mean are the same. The second is to remember that Thurstone (1935) (and Thomson somewhat less explicitly) developed these estimates as regression of $\mathbf{f}$ on $\mathbf{y}$, obtaining the expected value of $\mathbf{f}$ given $\mathbf{y}$, which is the mean or the EAP estimate.

4 Alternative Approaches to Scoring, and Controversy

4.1 Bias Adjustments

4.1.1 IRT Score Estimates

One area in which different challenges have presented themselves for the use of IRT scores compared to factor scores is the need for bias corrections. While the term bias has several uses that may reasonably be applied to test scores, we are concerned here with any systematic difference between the values of score estimates and the underlying latent scores, defined as the (possibly conditional) average difference between the true value and the estimate.[9]

Bartlett's factor score estimates are conditionally unbiased (Bartlett, 1938); Thomson-Thurstone regression estimates are biased inward toward the mean, as is always the case with regressed estimates that minimize the sum of the squared discrepancy between the estimates and the true values (Thomson, 1938).

The situation is only one third so simple for IRT scores. EAP estimates minimize the total squared difference between the estimates and the true values, and in so doing shrink toward the population mean like regressed estimates, so they are biased. MAP estimates approximate EAP estimates, but unlike the case with linear normal factor analysis, the posterior densities are not normal, or even symmetrical, so the MAP estimate is not exactly equal to the EAP estimate. IRT MAP estimates are neither unbiased nor the values that minimize squared loss. Most unlike the case for factor analysis, IRT ML estimates are not unbiased. Indeed, they are usually technically infinitely biased, because for any value of θ there is some non-zero probability that a perfect response pattern (all 1s or all 0s for dichotomous items) will be observed. Such perfect response patterns are associated with infinite (positive or negative) ML estimates. Infinite estimates included in a mean, with any non-zero weight, makes the average and bias infinite except in a perfectly symmetric case where the infinities cancel. In practice, computer implementations of

[9] By contrast, Kleinbort et al. (2022) discuss bias in factor models in the sense of group differences which are represented in multiple observed values, encoded in the factor model, represented in factor scores, and propagated to values imputed, via the factor model, where observations were originally missing. In this case, the score estimates may be unbiased estimates (in our sense) of the underlying latent scores, where those constructs are defined so as to include the difference between groups.

IRT place some positive and negative limits on the values of ML estimates for perfect response patterns, or make them missing, so simulation studies report finite bias for ML estimates. But that is artificial.

Weighted likelihood (WL) estimates were devised by Warm (1989) to correct the bias in ML estimates. In practice, the WL estimates fall between the inwardly-biased (toward the mean) EAP or MAP estimates and the outwardly-biased (toward the extremes) ML estimates. WL estimates are obtained by finding the maximum over θ of

$$L_w(\theta|\mathbf{u}_i) \propto \prod_j T(u_{ij}|\theta)w(\theta) \tag{37}$$

which is equation 14 with $w(\theta)$, a weight function, added, or 15, with a weight function substituted for the population density $\phi(\theta)$. For standard two-parameter IRT models for dichotomous data, $w(\theta) = I(\theta)^{1/2}$, the square root of the Fisher information function for the test. Warm (1989) derived this result from the known bias of the ML estimate by choosing the function $w(\theta)$ to cancel the first-order bias.[10] WL estimates shrink or regress the ML estimates toward the mode of $w(\theta)$, but usually not as much as the MAP shrinks toward the mode of a normal population density.

The formulation of the estimation procedure as maximization of equation 37 also makes $w(\theta)$ appear to be like the population distribution, or a Bayesian prior density for θ, but Warm (1989) was at pains to write that he thought such a Bayesian interpretation of $w(\theta)$ was wrong. Warm pointed out that $w(\theta)$ is a function of θ and the item parameters, so it is the same for persons from any population who respond to the same test items, while a sensible Bayesian population distribution could be different for those different populations. Another way to see that $w(\theta)$ may be difficult to think of as a population distribution is to notice that it *is* a function of the items, so in a computerized adaptive test (CAT) on which respondents see different items, $w(\theta)$ would be different for each person, even though all persons may be from the same population. More recently, however, Magis and Raîche (2012) have shown that for the two-parameter logistic model WL estimation is exactly Bayes modal estimation with Jeffreys' (1939, 1946) prior, and that the two methods are closely related but not identical for the three-parameter logistic. Magis (2015) provides equivalent WL and Jeffreys modal estimates for broad classes of polytomous models.

Warm (1989) showed with simulation that WL estimates outperformed ML estimates in all respects and MAP estimates in many respects. But that has been with $w(\theta)$ computed for a test that was well centered at the average of θ, so it was effectively a thicker-tailed population distribution than the normal density. It is not clear how rapidly the performance of WL estimates might deteriorate if the item set did not match the θ density. On the other hand, if a CAT is effectively tailoring the item set administered to each person, one could think it reasonable to have something like a Bayesian prior for θ in the model that shrinks estimated θ toward the central value of item difficulty for that person; Shao et al. (2020) proposed such a scoring method.

WL estimates have rarely if ever been used in practice. They have not been widely implemented in software, and there is little demand. Part, if not all, of the reason is that they make little difference. Most testing programs rescale θ estimates onto some reporting scale, like the well known *College Board* 200-800 scale with a mean of 500 and a standard deviation of 100. Because EAP or MAP or WL or ML estimates for responses to a fixed test are all in the same order, with only slight differences in relative spacing (assuming the infinite end effects of ML have been managed), transformation onto the reporting scale would obscure any differences arising from the original estimation method.

In addition, the impact of the WL estimate correction on relative spacing is larger for short tests. Longer, more informative, and more precise tests, such as large testing programs are compelled by practical concerns to use anyway, are less affected. All of the shrinkage mechanisms that involve a multiplicative term in the likelihood have this property, although the degree of effect varies. Each item contributes information additively to the test information function; the population distribution in EAP or MAP estimation, which is introduced as a multiplicative term in the likelihood function, also adds a quantity of information (Thissen & Orlando, 2001). The information added by the population distribution can be described as "like having an extra item;" it is constant, unrelated to the length of the test. For a short test with few items, the population distribution contributes substantially; for a long test with many items and high precision, the population distribution contributes relatively little. The $w(\theta)$ term, also a multiplicative component of the likelihood, similarly adds information,

[10] *First order bias* is bias of order $O(n^{-1})$ for n items, which is to say bias that is proportional to the inverse of the number of items. Biases of order $O(n^{-2})$ and smaller are usually ignored.

but the relative increment, compared to what comes from the items, increases as the square root of the length of the test. It does not keep up with the contribution of the items.

The test length effect can also be looked at from the opposite perspective, in which the regression to the mean of an EAP estimate, or the correction from WL, is lessened as the test lengthens and the standard error of the estimate decreases.

4.1.2 Factor Score Estimates within Larger Models

For estimation of factor scores in the linear-normal model, it might appear that the fact that Bartlett factor score estimates are unbiased would mean that bias is not an issue, but that is not the case. While Bartlett factor score estimates may be unbiased, statistics involving the latent variables computed using those estimates are not, in general, unbiased estimates of the same estimands. For example, the correlations among multiple factors are not well estimated by the correlations among either Bartlett or Thomson-Thurstone regression estimates. There is a long history of "correlation preserving factor score prediction methods" (ten Berge et al., 1999, p. 311); examples include proposals by Anderson and Rubin (1956), Green (1969), McDonald (1981), and Krijnen et al. (1996).[11]

Researchers may find it desirable to estimate regression coefficients both among latent variables and between latent variables and exogenous variables in models that are too large or otherwise unsuited for structural equations modeling. Examples abound when factor score estimates are used as test scores in operational testing programs in employee selection or development: In the creation of composite scores to optimize (usually concurrent) local or role-specific validity, in longitudinal studies of predictive validity, and in evaluation of operational programs for fairness, it is important to use the *existing* factor score estimates for all purposes, rather than varying the model or scoring method to suit the research purpose.

In such cases the researchers may use factor score estimates as data, and adjustments that yield correct estimates of the correlations among the latent variables may be insufficient. Skrondal and Laake (2001, p. 564) wrote "most researchers appear to contend that factor score regression will produce biased and/or inconsistent estimates of the structural parameters (e.g., Bartholomew, 1981; Bollen, 1989 ...)." Skrondal and Laake suggested a simple but effective solution which is to use Thomson-Thurstone regression factor score estimates for the predictor variable(s) and Bartlett estimates for the response variable(s). Croon (2002) pointed out that the underlying problem is the unreliability of the factor score estimates, and proposed correction formulae for regression, Bartlett, or Anderson-Rubin estimates to produce scores that could be used in standard analyses to obtain better estimates of regression coefficients between latent variables. Hoshino and Bentler (2013) proposed an alternative algorithm, which is to compute Bartlett estimates for all of the latent variables involved, then the mean and covariance matrix among those estimates by standard methods, and fit the latent variable regression model(s) using generalized least squares with the mean and covariance matrix for the Bartlett estimates as input.

4.1.3 IRT Score Estimates within Larger Models

Problems with the use of IRT score estimates as variables in subsequent analyses have seen less attention, probably because reporting scale scores is often an end in itself in applied applications of IRT. However, the same problems exist with the use of IRT scores in regression or structural equations models (Hoijtink & Boomsma, 1996; Lu et al., 2005; Mislevy et al., 1992), with the same impact on their use in employment testing, and with fewer solutions than have been offered for factor score estimates. Variants of the proposals by Croon (2002), Hoshino and Bentler (2013), or even Skrondal and Laake (2001) may be effective.

However, IRT is the home of a different solution to a related, but different, set of challenges. Primary reporting for the National Assessment of Educational Progress (NAEP) involves descriptions of the distributions of proficiencies (the latent variables for educational achievement) for populations and subgroups. Means, variances, and quantiles are important. Due to shrinkage (for MAP or EAP) or excess variance (for ML estimates), computing those statistics with IRT point estimates does not yield the right answers. Mislevy et al. (1992) describe the use of so-called *plausible values* instead of IRT score estimates to compute the summary statistics.

[11] This line of research illustrates a contrast between the factor scores and IRT literatures, that the factor scores literature is often about the *properties* of the estimates, while the IRT literature is more often concerned with their *accuracy*. Examples of the properties of the estimates that have concerned factor analysts involve the fact that their variances, and covariances with each other as well as with exogenous variables, are not the same as those of the true underlying values; test theorists have tended to take those properties of test scores as facts of relatively little interest. An example of the consideration of the accuracy of alternative estimates in the IRT literature is the study by Wainer and Thissen (1987), which compared several potentially robust estimates with respect to their conditional bias and precision.

Plausible values are (multiple) random imputations drawn from the posterior distribution for $\boldsymbol{\theta}$, which may be multidimensional. Plausible values are treated in data analysis as though they were individual test scores, although they are not. There are complications: To account for error of estimation, analyses using plausible values are repeated several times with distinct sets of draws from the posterior, and then the disparate results are averaged, with their variation included as a component of the associated standard errors.

In principle, plausible values technology was intended to provide unbiased results from secondary regression analysis of NAEP scores as well. In practice, the issues are a little more complicated: First, the plausible values only provide unbiased regression results for predictor variables that are included in the original so-called *conditioning* part of the IRT model: demographic background and other variables that are used as linear predictors of the latent variable(s) in the scoring model (Mislevy et al., 1992; Schofield et al., 2015). Second, the bias-avoidance feature of plausible values only works when the latent variables are the dependent or response variables in the regression analyses; if they are among the independent or predictor variables, bias arises. Schofield et al. (2015) document this problem and provide a solution for structural equation modeling.

4.2 Normal Approximations in IRT

Iterative estimation of IRT ML, WL, or MAP score estimates, or numerical integration for EAP estimates, was a computational challenge when IRT began to be considered for operational use fifty years ago, and such scoring systems remain difficult to explain to users even now after the computing time required has become negligible. At times, normal approximations have been suggested to speed up or (apparently) simplify the computation of IRT score estimates.

Owen (1969, 1975) developed a Bayesian procedure for IRT scores for CATs at a time when computers were barely fast enough to compute the required score estimates after each item response in time for selection of the next item. Assuming the population distribution was normal, and the trace line model for the (first) item response was the normal ogive, Owen analytically derived closed-form equations for the mean (EAP estimate) and variance of the posterior given either a correct or incorrect response. Even in the early 1970s those values could be computed quickly. Then in the context of a CAT, Owen's procedure was to update the population, or prior, distribution to be the posterior distribution and administer the next item, using the same equations to compute the next posterior mean, and so on through the test. This algorithm repeatedly approximated the non-normal IRT posterior with a normal density, but the loss of precision was relatively small and the saving in computational time was large. Owen's Bayes estimates, as they were called, are rarely used (or even remembered) because computers are now fast enough to compute EAP estimates effectively instantly with no approximations. Nevertheless, the procedure remains an interesting footnote to the history of latent variable score estimates.

In this century, Thissen, Nelson, and Swygert (2001) used normal approximations to propose solutions to different problems. The first application they considered was an IRT-based combination rule for scoring educational achievement tests with a block of multiple-choice (or other) items scored correct or incorrect, and a second block of free response items polytomously scored. While one *can* compute response-pattern IRT scores for such tests, those are difficult to explain to broad audiences of consumers of educational test scores. Thissen, Nelson, and Swygert (2001) used normal approximations to the IRT posterior densities for summed scores for the dichotomously-scored part and the polytomously-scored part, then combined the two part-scores. The result was a kind of weighted combination of the two parts, but based on IRT analysis instead of arbitrary values, and with weights that dynamically depend on the relative precision of evidence from the two parts of the test at each combined-score level. While those weighted scores have been used only rarely in practice, they serve to describe the relative weight IRT attributes to the two parts of such a test, as a standard for comparison with the more commonly used arbitrary weighting schemes.

Thissen, Nelson, and Swygert (2001) extended the procedure to provide approximate IRT scores for patterns of summed scores over more than two parts of a test. They illustrated this extension with an application to a testlet-based CAT, in which examinees respond to several adaptively-selected fixed blocks each comprising a handful of items; the approximation to response-pattern scoring was very close. While full response pattern scoring is challenging to explain in the large-scale educational context, a system that yields a score that appears to be a weighted combination of scores derived from a few blocks of items may appear to be more along the lines of what

is expected as a test score. Green (2002) examined the performance of simpler fixed-weight scoring systems for CATs, again to simplify explanation of the scoring systems. Green's scores were various averages using weights based on the item difficulty parameters. Green acknowledged that IRT scoring effectively used weights that vary as a function of proficiency, as do the relative weights in the proposal by Thissen, Nelson, and Swygert (2001); scores computed with difficulty weights alone were somewhat less precise than the gold standard IRT scores. Nevertheless, in explanations of CAT scoring it is sometimes useful to make use of the idea that "a CAT score is essentially a system for giving more credit for more difficult questions answered correctly, coupled with some credit for the difficulty of items answered incorrectly" (Green, 2002, p. 18).

In an entirely different context, the plausible values described in the previous section as random imputations from the IRT posterior density are actually, in practice, random numbers drawn from normal distributions with mean and variance computed as IRT score estimates and their associated error variances (Mislevy et al., 1992). This is done because drawing random normal deviates is a well-understood problem with existing efficient software implementations, whereas sampling from the actual IRT posterior would be much more computationally intensive at best. The IRT posterior panel in Figure 1 illustrate that such densities are often bell-shaped and normal-looking.

4.3 "Factor Score Indeterminacy"

The fact that there is no conceptual difference between factor score estimation and IRT score estimation raises the question: Why is there is an enormous literature full of controversy on the subject of "factor score indeterminacy," but no parallel IRT literature?[12]

The answer lies in different conceptions of the latent variable by different authors at different times. Maraun (1996b, p. 517) suggests the distinction between "The *alternative solution* position" that "considers the latent common factor to be a random variate whose properties are determined by functional constraints inherent in the model" and the "*posterior moment* position" that "considers the latent common factor to be a single random entity with a non-point posterior distribution, given the manifest variables." The former position leads to the conclusion that the latent variable is ill-defined in the sense that there are infinitely many alternative solutions, and the conclusion that factors scores are indeterminate and (in extreme statements) not useful. The posterior moment position produces the factor score estimation procedures described in previous sections, and is essentially the only perspective in the IRT literature.

Guttman (1955, p. 65) provided a succinct statement of the alternative solution position in the introduction of the article that reignited concern with factor score indeterminacy in the 1950s: "If ξ is the observed score matrix, a *direct* analysis would consist in finding factor score matrices η and ζ such that

$$\boldsymbol{\xi} = \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\zeta} \tag{38}$$

where $\mathbf{A}$ is some real matrix of common-factor loadings." Gutmann's equation 1 (numbered 38 here) is our equation 1 with different notation: $\boldsymbol{\xi}$ for $\mathbf{y}_i$, $\mathbf{A}$ for $\boldsymbol{\Lambda}$, $\boldsymbol{\eta}$ for $\mathbf{f}_i$, and $\boldsymbol{\zeta}$ for $\boldsymbol{\varepsilon}_i$. Stating the model in this way suggests that the goal is to simultaneously fill in values for the common factor scores (Guttman's $\boldsymbol{\eta}$ or our $\mathbf{f}_i$) and the error or unique or specific factor scores (Guttman's $\boldsymbol{\zeta}$ or our $\boldsymbol{\varepsilon}_i$), The oft-repeated point of factor score indeterminacy is that when the problem is set up that way, there are more unknowns than there are equations. For any values of the common factor scores ($\boldsymbol{\eta}$ or our $\mathbf{f}_i$), corresponding values of the error terms ($\boldsymbol{\zeta}$ or our $\boldsymbol{\varepsilon}_i$) can be obtained by subtraction. This is not a new argument; Wilson (1928) took this position in immediate response to the publication of Spearman's (1927) book. Confusingly, Thomson (1935, 1936) illustrated the alternative solution position in the same articles that introduced his procedure for computing point estimates that represent the posterior moment position. Others reiterated the alternative solutions position before the publication of Guttman's (1955) article, but the extremity of positions taken is illustrated by

[12] This section is about *factor score indeterminacy* (see, e.g., Grice, 2001). There are several other aspects of factor analysis or multidimensional IRT models to which the term *indeterminacy* may be applied: Principal among those is *rotational indeterminacy* for multiple factor models (Harmon, 1976; Mulaik, 1972); we avoid that subject here by specifying that we consider scores only for dimensionally-identified confirmatory factor analysis models. There is always indeterminacy of *reflection* in dimensional latent variable models; the positive-negative direction of the scores is arbitrarily determined in some way. There are also potential data-dependent indeterminacies of various kinds: The likelihood surface for the parameters of the multidimensional factor or IRT model may be multimodal, leading to indeterminacy of maximum likelihood solutions (Rubin & Thayer, 1982, 1983; Li Cai [personal communication, August 22, 2022]). The 3PL IRT model, and probably other IRT models with additive guessing components, may have a bimodal likelihood for response pattern scores (Samejima, 1973; Yen et al., 1991). None of these aspects of the models are the subject of this section.

Guttman's (1955, p. 79) conclusion "If more direct observations on the $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ cannot be made than statistical analysis of $\mathbf{R}$ and $\boldsymbol{\xi}$, the Spearman-Thurstone approach may have to be discarded for lack of determinacy of its factor scores." Subsequent publications have emphasized that factor score indeterminacy is a threat to the usefulness of factor analysis (Schönemann & Wang, 1972; Steiger, 1979, 1994, 1996a, 1996b; Steiger & Schonemann, 1978), or have reproduced the alternate solutions position uncritically along with discussions of factor analysis as a method, and the computation of prediction or regression scores (Mulaik, 1972).

Maraun (1996b, p. 517) concludes that "(a) The issue of indeterminacy centres on the criterion for the claim "X is a latent common factor to Y"; (b) the alternative solution position is correct, the posterior moment position representing a conflation of the criterion, which is provided by the equations of the model, with metaphors, analogies, and senses of "factor" that are external to the model." We agree with (a) and disagree with the first clause of (b), because a statistical model is nothing if not a mathematical expression of metaphor or analogy, drawing on our knowledge of the world external to the model.

While McDonald (1974) and Bartholomew (1981) provide descriptions of the posterior moment position in the factor analysis literature, to be as clear as possible about its relation with IRT, we rely on the description of the model in previous sections: To quickly recapitulate, we express the multiple factor model as equation 1. Then the likelihood for the data as a function of the factor scores (assuming $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ are known) is equation 27, above.

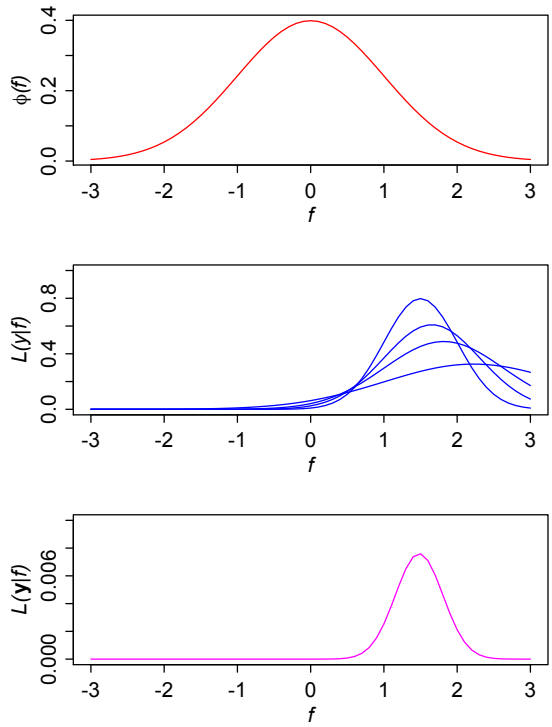
We make two observations about writing the factor analysis model using equation 27: (1) This presents the model in direct parallel with an IRT model, as a probability (likelihood) statement about the observed data, and (2) this eschews the *appearance* of a linear equation into which one could insert arbitrary values of $\mathbf{f}_i$ and corresponding values of $\boldsymbol{\varepsilon}_i$, thus inviting the alternate solutions story. As is the case with any statements of IRT models for categorical data, equation 27 simply has no place to put the alternate solutions. So, if the model had been written as likelihood like this from the beginning, the entire factor score indeterminacy saga would not have arisen.

As we pointed out in a previous section, if a standard normal population distribution for $\mathbf{f}$ is included in the model, with correlation matrix $\boldsymbol{\Phi}$ among the factors, Bayes theorem can be used to provide the likelihood of the factor scores given the data, and the log likelihood. That is equation 29 above. The maxima of equations 27 and 29 are the ML or MAP/EAP estimates of $\mathbf{f}_i$. This makes salient the problem with lists of alternate solutions: There are, for any set of data, alternative values of the factor scores and errors that correspond perfectly with the data in the sense of equation 1. But the *likelihoods* of those alternative values are very different. If corresponding likelihoods were also reported in the lists of alternate solutions, it would have been clear that all of the alternative values of $\mathbf{f}_i$ have lower likelihood than the MAP (Thomson-Thurstone regression) estimates. For example, Thomson (1935) illustrated the alternative solutions idea with hypothetical data and factor score values of 1.00 and 1.96 which appear to yield the same fit when inserted into the model algebra; he also wrote that the regression estimate of the factor score for the example is 1.48. This was intended to suggest that there is some sense in which the disparate values 1.00 and 1.96 could be equally good as factor scores. But no mention is made of the fact that the likelihood at $f = 1.48$ (the MAP/EAP estimate) is approximately three times higher than at either $f = 1.00$ or $f = 1.96$. Considering likelihood, one chooses the Thomson-Thurstone regression estimate. The lower panel of Figure 2 shows the posterior likelihood for the data Thomson used; the mode is at $f = 1.48$, and the likelihood is much lower at the values $f = 1.00$ and $f = 1.96$.

To conclude, IRT model equations have no place to put alternative solutions, so indeterminacy has never arisen in test theory.[13] If the factor analysis model is written like an IRT model, as a probability or likelihood model, there is no place for alternative solutions and the whole issue does not arise. And what is omitted from the alternative solutions argument is the fact that the alternative solutions have varying likelihood, and the standard statistical procedure under those circumstances would be to select a solution

[13] Yang Liu (personal communication, August 9, 2022) has observed that it is possible to express an IRT model using a data generating equation rather than in likelihood terms. For example, the logistic model of equation 13 could be $u_{ij} = \mathbb{1}\{A_{ij} \leq [-a_j(\theta_i - b_j)]\}$, using the indicator function for person i and item j with $\theta_i \sim N(0,1)$ and $A_{ij} \sim \text{logistic}(0,1)$. One could then imagine indeterminacy of θ_i jointly with A_{ij}, by analogy with the factor analytic argument. This data-generating form of the IRT model is closely related to that described by Lord and Novick (1968, pp. 370-371) for the normal ogive model, McCullagh and Nelder's (1989, pp. 151-155) generalized linear model for ordinal data, and the basis for the data augmentation in Albert's (1992) MCMC estimation algorithm, again for the normal ogive IRT model. A difference is that in the notation of the latter three examples, the person-item specific random variable is a function of θ while A_{ij}, is independent of θ, corresponding to the unique term in the factor model and the core element in the alternative solutions argument.

Figure 2



Upper panel: The $N(0,1)$ population distribution $\phi(f)$. Center panel: Four likelihoods $L(y_{ij}|f_i)$ for four values of $\mathbf{y} = [1.3416, 1.3844, 1.4071, 1.4071]$ and $\boldsymbol{\lambda} = [0.8944, 0.8367, 0.7746, 0.6325]$ from Thomson's (1935, pp. 247-250) example. Lower panel: The posterior, $L_p(f|\mathbf{y}_i)$ from equation 19, which is the product of the likelihoods in the center panel times the red population distribution in the upper panel.

with maximum posterior likelihood, or an average, which, for a linear-normal model, are the same thing.

None of what is argued in this section is new. McDonald (1996, p. 596) also appealed to the correspondence between IRT and factor score estimates when he wrote in response to Maraun (1996b) that the alternative solutions position "can be taken as saying that because latent variables are random they cannot be estimated. But this is patently untrue. The ultimate goal of item response theory is to use the item parameters from a calibration sample to estimate the latent traits of one or more examinees (possibly including all those in the calibration sample) and to estimate the error of the psychological measurement. . . . in the linear case the model reduces to the common factor model and the estimates to Bartlett's ML/GLS expression." But in another of the responses comprising the back-and-forth between Maraun and several commentators, Bartholomew (1996, p. 631) wrote "Dr. Maraun's (1996a) response serves to emphasize the mutual incomprehension which has characterized the debate over the years." No doubt mutual incomprehension will continue.

5 Conclusion

We have sketched the historical development of latent variable scoring strategies in the IRT and factor analysis literatures, observing that the most commonly used score estimates in both traditions are fundamentally the same. Methods of calculation differ: IRT score estimates require iterative computation or numerical integration, while factor score estimates derived from the same likelihood principles are amenable to closed-form solutions. That, and the difference in the ways models have been expressed in the two traditions, has made the estimates look very different to students, at least up until the publication of recent integrations (Bartholomew et al., 2011; Skrondal & Rabe-Hesketh, 2004). And it is the case that the two traditions can be merged: Thissen (1989) described a system to estimate *skeletal maturity* that combined categorical and continuous indicators, with logistic IRT models for the former and the linear-normal model for the latter, using the IRT approach to ML score estimation. While the methods used by Thissen (1989) were not entirely latent-variable based, Nahhas et al. (2013) has suggested that system be updated using contemporary latent variable approaches.

In applications of methods from both traditions, it is sometimes desirable to compute score estimates in the presence of individual item responses or variable measurements that may be missing at random. That can be done with any of the procedures described in the preceding sections. For IRT-like likelihood-based computations that have no closed form solution, one simply omits the terms in the log likelihood associated with the missing observation(s); calculation remains the same. Comparably, when using a modern linear equation solver to produce Thomson-Thurstone factor score estimates, one simply omits the $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}'$, and $\boldsymbol{\Theta}^{-1}$ matrix rows and columns corresponding to the missing observations, and proceeds with the same calculation. If one is using the apparently-simpler closed-form factor score solutions with regression coefficients, missing observations may require recomputation of the regression coefficients for the non-missing subset of variables; that is more calculation, not less, but at modern computer speeds still negligible. And on the factor analysis side, there is a shortcut for some simple models: those in which each observed variable

is associated with only one factor. The simplest special case is the unidimensional model, for which the estimates can be computed as weighted sums of the non-missing observations (Thissen & Thissen-Roe, 2020).

Due to differences in the context of score usage, challenges have led to different solutions in the IRT and factor analytic traditions. While Bartlett factor score estimates are unbiased, regression- or SEM-model coefficients using any factor score estimates as data are biased, and several solutions have been offered (Croon, 2002; Hoshino & Bentler, 2013; Skrondal & Rabe-Hesketh, 2004). While the same problems can arise with IRT score estimates, the situation does not arise in the most common use cases so it has received less attention. On the other hand, IRT has been the home of *plausible values* technology to avoid bias in the estimation of some population statistics (Mislevy et al., 1992).

The linear normal factor model is blessed with highly tractable Gaussian likelihoods and posteriors; IRT likelihoods and posterior densities are not Gaussian, but normal approximations have been used to good effect in some special cases to obtain some of the computational simplicity enjoyed with the factor model (Mislevy et al., 1992; Owen, 1969, 1975; Thissen, Nelson, & Swygert, 2001).

Finally, factor analysis alone has been the home of decades of controversy over *factor score indeterminacy*. That is an artifact of history and the ways the models have been written in the IRT and factor analytic literatures. The factor analysis model was (and still is) often written as an algebraic expression with symbols for both the underlying latent variables and the error components, extending an implicit invitation to fill in both with numeric values. IRT models have been written as probabilistic statements about the way the observed data depend on the latent variables(s); those statements have no place to put numeric estimates of the error terms. So IRT has never been plagued with questions of indeterminacy, and the scores are simply used as what all test scores are: estimates with error.

We hope that integration of the IRT and factor analytic traditions serves to make both easier to understand.

Acknowledgments

We thank Li Cai, Yang Liu, Alberto Maydeu-Olivares, and Jolynn Pek for insightful and helpful comments on an earlier draft. Any errors that remain are our own.

References

Albert, J. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269. https://doi.org/10.2307/1165149

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium in mathematical statistics and probability* (pp. 111–150). University of California Press. https://doi.org/10.1007/BF02289543

Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology*, *34*, 93–99. https://doi.org/10.1111/j.2044-8317.1981.tb00620.x

Bartholomew, D. J. (1996). Response to Dr. Maraun's first reply to discussion of his paper. *Multivariate Behavioral Research*, *31*, 631–636. https://doi.org/10.1207/s15327906mbr3104_15

Bartholomew, D. J., Deary, I., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology*, *62*, 569–582. https://doi.org/10.1348/000711008x365676

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons. https://doi.org/10.1002/9781119970583

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, *28*, 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x

Bartlett, M. S. (1938). Methods of estimating mental factors. *Nature*, *141*, 609–610. https://doi.org/10.1038/141246a0

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* (pp. 395–479). Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. https://doi.org/10.1177/014662168200600405

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–223). Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781410602961-16

Devlieger, I., Mayer, A., & Rosseel, Y. (2015). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*, 741–770. https://doi.org/10.1177/0013164415607618

Estabrook, R., & Neale, M. (2013). A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. *Multivariate Behavioral Research*, *48*, 1–27. https://doi.org/10.1080/00273171.2012.730072

Ferguson, G. A. (1942). Item selection by the constant prcocess. *Psychometrika*, *7*, 19–29. https://doi.org/10.1007/bf02288601

Green, B. F. (1969). Best linear composites with a specified structure. *Psychometrika*, *34*, 301–318. https://doi.org/10.1007/BF02289359

Green, B. F. (2002). *Fixed-weight methods of scoring computer-based adaptive tests* (LSAC Research Report Series No. 97-12). Law School Admission Council.

Grice, J. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450. https://doi.org/10.1037/1082-989x.6.4.430

Guttman, L. (1940). Multiple rectilinear prediction and the resolution into components. *Psychometrika*, *5*, 75–99. https://doi.org/10.1007/bf02287866

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *The British Journal of Statistical Psychology*, *8*, 65–81. https://doi.org/10.1111/j.2044-8317.1955.tb00321.x

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the doise is subjext to error* (Technical Report No. 15). Applied Mathematics and Statistics Laboratory, Stanford University.

Harmon, H. (1976). *Modern factor analysis* (Third ed.). University of Chicago Press.

Hoijtink, H., & Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika*, *61*, 313–330. https://doi.org/10.1007/bf02294342

Hoshino, T., & Bentler, P. M. (2013). Bias in factor score regression and a simple solution. In A. R. de Leon & K. C. Chough (Eds.), *Analysis of mixed data: Methods & applications* (pp. 43–61). Chapman and Hall/CRC. https://doi.org/10.1201/b14571-5

Jeffreys, H. (1939). *Theory of probability*. Oxford University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *186*, 453–461. https://doi.org/10.1098/rspa.1946.0056

Kelley, T. L. (1927). *The interpretation of educational measurements*. World Book.

Kleinbort, A., Thissen-Roe, A., Chakraborty, R., & Szary, J. (2022). *Considerations in group differences in missing values*. Presentation at the International Meeting of the Psychometric Society, Bologna, Italy, July 11-15.

Krijnen, W. P., Wansbeek, T., & ten Berge, J. M. (1996). Best linear predictors for factor scores. *Communications in Statistics - Theory and Methods*, *25*, 3013–3025. https://doi.org/10.1080/03610929608831883

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings fo the Royal Society of Edinburgh*, *60*, 64–82. https://doi.org/10.1017/S037016460002006X

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings fo the Royal Society of Edinburgh*, *61-A*, 273–287. https://doi.org/10.1017/s0080454100006282

Lazarsfeld, P. F. (1950a). The interpretation and computation of some latent structures. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 413–472). Wiley. https://doi.org/10.2307/2571672

Lazarsfeld, P. F. (1950b). 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* (pp. 362–412). Wiley. https://doi.org/10.2307/2571672

Ledermann, W. (1939). On a shortened method of estimation of mental factors by regression. *Psychometrika*, *4*, 109–116. https://doi.org/10.1007/bf02288490

Loncke, J., Eichelsheim, V., Branje, S., Buysse, A., Meeus, W., & Loeys, T. (2018). Factor score regression with social relations model components: A case study exploring antecedents and consequences of perceived support in families. *Frontiers in Psychology*, *9:1699*, 1–19. https://doi.org/10.3389/fpsyg.2018.01699

Lord, F. M. (1952). *A theory of test scores*. (Psychometric Monograph No. 7). Psychomtric Corporation. Retrieved October 7, 2021, from http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, *18*, 181–194. https://doi.org/10.1007/bf02289028

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

Lu, I. R. R., Thomas, R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on irt scores. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 263–277. https://doi.org/10.1207/s15328007sem1202_5

Magis, D. (2015). A note on weighted likelihood and Jeffreys modal estimation of proficiency levels in polytomous item response models. *Psychometrika*, *80*, 200–204. https://doi.org/10.1007/s11336-013-9378-5

Magis, D., & Raîche, G. (2012). On the relationships between Jeffreys modal and weighted likelihood estimation of ability under logistic IRT models. *Psychometrika*, *77*, 163–169. https://doi.org/10.1007/S11336-011-9233-5

Maraun, M. D. (1996a). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research*, *31*, 603–616. https://doi.org/10.1207/s15327906mbr3104_13

Maraun, M. D. (1996b). Metaphor taken as math: Indeterminancy in the factor analysis model. *Multivariate Behavioral Research*, *31*, 517–538. https://doi.org/10.1207/s15327906mbr3104_6

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate analysis*. Academic Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. CRC Press LLC. https://doi.org/10.1007/978-1-4899-3242-6

McDonald, R. P. (1974). The measurement of factor indeterminacy. *Psychometrika*, *39*, 203–222. https://doi.org/10.1007/bf02291469

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika*, *46*, 337–341. https://doi.org/10.1007/BF02293740

McDonald, R. P. (1996). Latent traits and the possibility of motion. *Multivariate Behavioral Research*, *31*, 593–602. https://doi.org/10.1207/s15327906mbr3104_12

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*, 511–536. https://doi.org/10.1007/S11336-011-9223-7

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154. https://doi.org/10.3102/10769986017002131

Mulaik, S. A. (1972). *The foundations of factor analysis*. McGraw-Hill. https://doi.org/10.1201/b15851

Nahhas, R. W., Sherwood, R. J., Chumlea, W. C., & Duren, D. L. (2013). An update of the statistical methods underlying the FELS method of skeletal maturity assessment. *Annals of Human Biology*, *40*, 505–514. https://doi.org/10.3109/03014460.2013.806591

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Bulletin No. RB-69-92). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1969.tb00772.x

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–256. https://doi.org/10.1080/01621459.1975.10479871

Rubin, D., & Thayer, D. (1982). EM algorithms for ML factor analysis. *Psychometrika*, *47*, 69–76. https://doi.org/10.1007/bf02293851

Rubin, D., & Thayer, D. (1983). More on EM for ML factor analysis. *Psychometrika*, *48*, 253–257. https://doi.org/10.1007/bf02294020

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, *38*, 221–233. https://doi.org/10.1007/bf02291115

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables

with covariates. *Psychometrika*, *80*, 727–747. https://doi.org/10.1007/s11336-014-9415-z

Schönemann, P., & Wang, M. (1972). Some new results on factor indeterminacy. *Psychometrika*, *37*, 61–91. https://doi.org/10.1007/bf02291413

Shao, C., Thissen, D., Cai, L., Cappaert, K., Edwards, M. C., & Shen, Y. (2020). *Proficiency estimation in computerized adaptive testing using a locally objective prior*. Presentation at the virtual annual meeting of the National Council on Measurement in Education, Sept. 10.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*, 563–576. https://doi.org/10.1007/bf02296196

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall–CRC. https://doi.org/10.1201/9780203489437

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. Macmillan.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika*, *44*, 157–167. https://doi.org/10.1007/bf02293967

Steiger, J. H. (1994). Factor analysis in the 1980's and the 1990's: Some old debates and some new developments. In I. Borg & P. Mohler (Eds.), *Trends and perspectives in empirical social research* (pp. 201–224). DeGruyter. https://doi.org/10.1515/9783110887617.201

Steiger, J. H. (1996a). Coming full circle in the history of factor indeterminacy. *Multivariate Behavioral Research*, *31*, 617–630. https://doi.org/10.1207/s15327906mbr3104_14

Steiger, J. H. (1996b). Dispelling some myths about factor indeterminacy. *Multivariate Behavioral Research*, *31*, 539–550. https://doi.org/10.1207/s15327906mbr3104_7

Steiger, J. H., & Schonemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 136–178). Jossey-Bass.

ten Berge, J. M., Krijnen, W. P., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications*, *289*, 311–318. https://doi.org/10.1016/S0024-3795(97)10007-6

Thissen, D. (1989). Statistical estimation of skeletal maturity. *American Journal of Human Biology*, *1*, 185–192. https://doi.org/10.1002/ajhb.1310010207

Thissen, D., & Thissen-Roe, A. (2020). Factor score estimation from the perspective of item response theory. In M. Wiberg, D. Molenaar, J. González, U. Bockenhold, & J.-S. Kim (Eds.), *Quantitative Psychology: 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019* (pp. 171–184). Springer.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141–186). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410604729-9

Thissen, D., Nelson, L., & Swygert, K. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—approximation methods for scale scores. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 293–341). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410604729-15

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410604729-8

Thomson, G. H. (1935). The definition and measurement of "*g*" (general intelligence). *The Journal of Educational Psychology*, *26*, 241–262. https://doi.org/10.1037/h0059873

Thomson, G. H. (1936). Some points of mathematical technique in the factorial analysis of ability. *Journal of Educational Psychology*, *27*, 36–54. https://doi.org/10.1037/h0062007

Thomson, G. H. (1938). Methods of estimating factor scores. *Nature*, *141*, 246. https://doi.org/10.1038/141246a0

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

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

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*,

12, 339–368. https : / / doi . org / 10 . 3102 / 10769986012004339

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/bf02294627

Wilson, E. B. (1928). On hierarchical correlation systems. *Proceedings of the National Academy of Science*, *14*, 283–291. https://doi.org/10.1073/pnas.14.3.283

Woodbury matrix identity. (2021). Retrieved May 7, 2021, from https : / / en . wikipedia . org / wiki / Woodbury_matrix_identity

Yen, W. M., Burket, G., & Sykes, R. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika*, *56*, 39–54. https://doi.org/10.1007/bf02294584