Article

Educational and Psychological Measurement
2016, Vol. 76(2) 325–338
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164415576958
epm.sagepub.com
SAGE

On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back

Tenko Raykov[1] and George A. Marcoulides[2]

Abstract

The frequently neglected and often misunderstood relationship between classical test theory and item response theory is discussed for the unidimensional case with binary measures and no guessing. It is pointed out that popular item response models can be directly obtained from classical test theory-based models by accounting for the discrete nature of the observed items. Two distinct observational equivalence approaches are outlined that render the item response models from corresponding classical test theory-based models, and can each be used to obtain the former from the latter models. Similarly, classical test theory models can be furnished using the reverse application of either of those approaches from corresponding item response models.

Keywords

binary item, classical test theory, item response theory, observational equivalence, unidimensionality

For much of the past century, classical test theory (CTT) was the dominant framework for developing multicomponent measuring instruments in the educational, behavioral, and social sciences. Part of its attraction was the simplicity of its fundamental

[1] Michigan State University, East Lansing, MI, USA
[2] University of California, Santa Barbara, CA, USA

Corresponding Author:
Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443A Erickson Hall, East Lansing, MI 48824, USA.
Email: raykov@msu.edu

observed score decomposition, which provided a very useful way of thinking about underlying constructs of substantive relevance for scholars involved in measurement. In the second half of the past century, interest alternatively in item response theory (IRT, item response modeling) increased substantially, eventually turning IRT into a major working methodology in test and scale construction, particularly during the past few decades.

An unfortunate by-product of this enhanced attention to IRT-based approaches for instrument development was a considerable degree of unjustified criticism of the general CTT framework. That criticism was to a certain extent a consequence of prior, widespread use of CTT-related procedures that were not correctly employed by some empirical scientists in the earlier part of the 20th century, during an era characterized by lack of sufficiently sophisticated statistical and methodological means allowing proper application of CTT for the purposes of test and scale construction. A result of that development was the tendency, especially in the second half of the past century, to neglect CTT as a general methodology informing about various aspects of instrument development and related issues. This tendency found various forms of expression and dissemination, both in writing and instruction, leading to what may be presently viewed as a preconceived and misleading notion of general deficiency of CTT when it comes to developing new or improving existing instruments, especially in the educational and psychological disciplines.

This deficiency notion with respect to CTT is not justified and in fact not correct. In reality there are strong relationships between CTT on one hand, when properly used, and IRT on the other hand. These relationships can be seen as closely tied to those between factor analysis (FA) and IRT, which have been pointed out in the methodological literature over the past 30 years or so (e.g., Kamata & Bauer, 2008; B. O. Muthén, Kao, & Burstein, 1991; Takane & de Leeuw, 1987; see also Kohli, Koran, & Henn, 2014). The CTT-IRT relationships, when appropriately highlighted and methodically clarified for empirical educational, behavioral, and social researchers, can in our view significantly contribute to improvements in their measurement related work as well as to tangible progress in the entire field of measurement. This clarification seems also to be needed because of the fact that the above methodological literature on the FA-IRT connection was largely developed within a fairly general framework applicable in multidimensional settings with ordinal items, and in addition was to a substantial degree of a technical nature. As a result, unfortunately it remained to a large extent inaccessible and abstruse for empirical scientists in these and cognate disciplines for the past several decades.

The goal of the present note is to revisit in light of that literature the relationship between CTT and IRT within a particular setting that is widely used currently in measurement contexts in educational and psychological research. By doing so, we hope to clarify some apparent misconceptions and imprecise beliefs about the supposed deficiencies of CTT relative to IRT. The remainder of this discussion is specifically concerned with the equivalence between CTT-based modeling and item response modeling for unidimensional tests or scales consisting of binary or binary scored measures with no guessing. The following discussion also provides a useful setup to underscore the close connections between CTT and IRT that in our opinion need to be highlighted and brought to the attention of empirical scientists.

Classical Test Theory Concepts for Homogeneous Instruments With Binary Items

The well-known CTT equation for the observed score on a given measure, denoted X_j, from a multicomponent instrument consisting of the manifest variables X_1, X_2, . . ., X_p ($p > 0$), is

$$X_j = T_j + E_j, \tag{1}$$

where T_j and E_j are its corresponding true and error scores (e.g., Zimmerman, 1975; $j = 1, \ldots, p$; in the remainder, we suppress the individual subscript for simplicity of notation). For many years, especially in the first part of the 20th century, it was widely held that the true score could only be defined when X_j is a measure on an interval or ratio scale. This incorrect view hampered progress in the measurement field in the educational, behavioral, and social sciences for a number of decades. (We point out that the CTT decomposition 1 of observed score into the sum of true and error score is valid for any given measure, regardless of whether it is a part of a multicomponent instrument or considered separately from any other measure; see next.)

Existence of the True Score and Error Score Construction

The observed score in Equation (1), for any prespecified individual, is a random variable pertaining to the administration of the jth measure to him or her. In particular, the mean of X_j, denoted $\mathcal{E}(X_j)$ and equal by definition to the associated true score T_j (Zimmerman, 1975), will exist as long as its variance, denoted $Var(X_j)$, exists, that is, as long as $Var(X_j) < \infty$ (e.g., Apostol, 2013; $j = 1, \ldots, p$). The last sufficient condition of finite variance, and hence the implied existence of the mean $\mathcal{E}(X_j) = T_j$, can be considered fulfilled practically in all empirically relevant cases (e.g., Lord & Novick, 1968). Therefore, for any studied person and a given measure, X_j, the existence of his or her true score on it, $T_j = \mathcal{E}(X_j)$, is ensured in all practically relevant cases regardless of the nature or scale of X_j. In particular, as long as the variance of X_j is finite ($1 \leq j \leq p$), T_j exists whether or not X_j is a binary, binary scored, or ordinal item, rather than only if this item is an interval or ratio scaled measure (as has been also incorrectly stated in several widely circulated sources over the past few decades). The error score associated with X_j is then defined by subtraction, namely, as $E_j = X_j - T_j$ ($j = 1, \ldots, p$).

Thus, if for the jth binary (or binary scored) item the two possible responses on it are denoted as follows:

$$X_j = \begin{cases} 1, & \text{if answer ``true,'' ``correct,'' or ``endorsing''} \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

then the CTT decomposition for that item is valid and of the following form:

$$X_j = T_j + E_j = \mathcal{E}(X_j) + E_j = P_j + E_j, \qquad (3)$$

with P_j being the probability of correct response on it ($j = 1, \ldots, p$; for simplicity in the remainder, we are referring to the response designated 1 in Equation 2 as ''correct'').

We present next two distinct observational equivalence approaches that can be used to develop, starting from appropriate CTT-based models, the popular one- and two-parameter logistic and normal ogive item response models. These approaches, although indirectly related, can be used independently of each other in furnishing the item response models from CTT-based models. As we will also point out, the reverse application of either of these approaches can be utilized in order to obtain any of these CTT-based models from a corresponding item response model.

From Classical Test Theory to Item Response Theory and Back: Approach 1

To demonstrate the CTT-IRT equivalence for unidimensional binary items with no guessing, which is the setting of concern in this article, we need to extend first the notion of congeneric tests to that setup.

A Congeneric Test Model for Binary Items

For a homogeneous multicomponent measuring instrument, the most general model relating its components within the CTT framework is the congeneric test model (CTM; e.g., Jöreskog, 1971). The CTM is quite popular in educational and psychological research with instrument components that are suspected of having a ''common genesis'', that is, measuring a common construct, and is readily testable using the latent variable modeling methodology (LVM; B. O. Muthén, 2002; see also Raykov & Marcoulides, 2011, for a testing approach). Part of the reason for the popularity of the standard CTM, is its assumption that the true scores pertaining to the observed measures are perfectly linearly related among themselves. Accordingly, in the notation used with Equation (1) above,

$$T_j = a_j + b_j T \qquad (4)$$

holds in this model, where T is the common true score for $T_1, \ldots, T_p$ (T could be taken, for instance, as T_1), while a_j and b_j are pertinent intercept and loading parameters ($j = 1, \ldots, p$). Equation (4) can also be seen as stipulating a deterministic relationship—that is, a relationship not containing an additional stochastic term—between any individual measure's true score and the common true score. (In general, a deterministic relationship need not be linear; see also below.)

With Equations (4) and (1), the standard CTM can be defined by the following set of observed variable equations:

$$X_j = a_j + b_j T + E_j, \tag{5}$$

where in addition $Var(T) = 1$ is assumed for model identification, and for the latter reason usually the error terms E_j are stipulated uncorrelated across measures as well, that is, the covariance matrix of the error score vector $\underline{E} = (E_1, \ldots, E_p)'$ is assumed diagonal (with underline denoting vector and prime denoting transposition; this error uncorrelatedness assumption is not needed in general, assuming model identification in an application, and is typically advanced for parsimony and convenience reasons as we will follow in the sequel; cf. Zimmerman, 1975).

An alternative representation of the standard CTM results from Equation (5) for a given common true score (see also Equation 3; $j = 1, \ldots, p$):

$$\varepsilon(X_j) = a_j + b_j T. \tag{6}$$

Since no restrictions are placed on the intercept and loading parameters, a_j and b_j, we notice from Equations (3) and (6) that for a binary or binary scored item its expectation—that is, the probability of correct response on it—is not restricted, whereas a probability must be bounded by 0 and 1. Hence, the standard CTM cannot be directly used or postulated in case of binary measures that are of interest in this article. However, employing the generalized linear modeling (GLIM) framework (e.g., Raykov & Marcoulides, 2011), the critical right-hand side of Equation (4) can be preserved if one considers instead the logit or probit of the expected observed score, that is, true score or the probability of correct response on the item, P_j:

$$ln\left[P_j/\left(1 - P_j\right)\right] = a_j + b_j T, \tag{7}$$

or

$$\Phi^{-1}\left(P_j\right) = a_j + b_j T, \tag{8}$$

respectively, where $ln(\cdot)$ denotes natural logarithm, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and for simplicity the same *a, b*, and *T* notation/symbols are used for the intercept, loading, and common construct in the right-hand sides of Equations (7) and (8) as in Equation (4) ($j = 1, \ldots, p$). (Strictly speaking, these symbols denote now different quantities but are used here for emphasizing the common structure of Equations 4, 7, and 8).

As can be observed from Equations (7) and (8), employing a particular link function such as the logit or probit there one relates within the GLIM framework a function of the response mean, rather than this mean itself, to one or more explanatory variables in their right-hand sides—in this case, the unobserved predictor *T*. In this way, the boundaries of 0 and 1 for the response probability (i.e., response mean) are no longer applicable since these functions of probability in their left-hand sides are no longer restricted. We should like to point out that Equations (7) and (8) represent deterministic nonlinear relationships between the probability of correct response on the *j*th measure—that is, its true score T_j in the present setting—and the common construct underlying all measures that is denoted *T*, with their right-hand sides having the same structure as that of Equation (4) in the model for congeneric measures, the CTM. Models (7) and (8) can be fitted to data using the LVM methodology (B. O. Muthén, 2002), employing for instance maximum likelihood or alternatively weighted least squares as methods for parameter estimation and model testing (L. K. Muthén & Muthén, 2014).

Because models (7) and (8) result from within the general setting of a congeneric test model, either of Equations (7) or (8) can be seen as defining a congeneric test model for binary items (CTMBI), correspondingly with the logit link or the probit link and no guessing. When in addition the constraint

$$b_1 = \cdots = b_p \tag{9}$$

holds, one could refer to the model defined by Equation (7) as an essentially tau-equivalent model with binary items based on the logit link, and to the model defined by Equation (8) as an essentially tau-equivalent model for binary items based on the probit link (cf., e.g., Raykov & Marcoulides, 2011).

We next observe that Equations (7) and (8) are in actual fact equivalent to those of the two-parameter logistic and two-parameter normal ogive item response models, respectively. This equivalence is readily obtained by employing: (i) the notation $T = \theta$, which is very popular within the IRT framework; (ii) b_j as the discrimination parameter of the *j*th item; and (iii) the ratio $(-a_j/b_j)$ as its difficulty parameter ($j = 1, \ldots, p$). The one-parameter logistic and one-parameter normal ogive models are then special cases of their corresponding two-parameter counterparts just mentioned, which result each when the restriction (9) holds in them.

This equivalence relationship can be utilized (a) to obtain a one- or two-parameter logistic or normal ogive item response model from the respective CTT-based model (7) or (8), correspondingly with or without the constraint (9), by using the earlier developments in this section; or alternatively (b) to furnish from a one- or two-parameter logistic or normal ogive item response model the corresponding CTT-based model (7) or (8) (or their special cases with constraint 9). That is, this observational equivalence relationship could be used to ‘‘move’’ from CTT to IRT, or alternatively from IRT to CTT in the setting of concern to this article—unidimensional binary items with no guessing. The equivalence also demonstrates how close these two frameworks in fact are, and may also be seen as a theoretical justification of the recent simulation-based results by Kohli et al. (2014) that show a lack of general numerical/estimation advantage of either framework over the other. (In actuality, the findings by Kohli et al., 2014, could be treated as illustrations of the CTT-IRT equivalence developments in this section; see also Conclusion section.)

From Classical Test Theory to Item Response Theory and Back: Approach 2

As an alternative approach for obtaining item response models from appropriate CTT-based models or conversely, one can use the following procedure based on an important assumption made when fitting latent variable models to data from discrete observed measures, which is attended to first.

The Underlying Normal Variable Assumption

An assumption that is usually advanced when analyzing categorical response variables in applied statistics, is that of a normal latent variable underlying an observed discrete variable, such as the above binary measure X_j (e.g., Agresti, 2002). It is this underlying variable, denoted X_j^*, which is of actual interest to measure, but because of serious measurement-related problems only its crude evaluation is possible in the associated discrete observed variable X_j (j = 1, . . ., p). This assumption has a long history of applications that are unrelated to CTT or IRT, and has been instrumentally used independently in genetics, attitude measurement, discrete choice, and economics, to name a few areas (e.g., Rabe-Hesketh & Skrondal, 2012); we will refer to it as the underlying normal variable (UNV) assumption. Accordingly, in relation to an associated unknown threshold τ_j the observed score on the measure of concern results as

$$X_j = \begin{cases} 1, & \text{if } X_j^* > \tau_j \\ 0, & \text{if } X_j^* \leq \tau_j \end{cases}. \tag{10}$$

(Note that because of the continuity of the random variable X_j^*, it is immaterial how X_j is defined when $X_j^* = \tau_j$, as long as it takes a finite value there; j = 1, . . ., p).

Classical Test Theory Decomposition Associated With Binary or Binary Scored Measures

When all items in a given multicomponent measuring instrument are binary or binary scored (with no guessing), the UNV assumption can be advanced with respect to each one of them, as is usually done in applications (e.g., Raykov & Marcoulides, 2011). Following this assumption, for each binary measure X_j there is an associated underlying normal variable X_j^* and threshold τ_j, such that Equation (10) holds (j = 1, . . ., p).

Considering now the set of p underlying normal variables X_j^* (j = 1, . . ., p), one easily realizes that they themselves are random variables with individual realizations for each studied subject (in a given sample or population of interest), which are not observed. In practically all relevant empirical cases, one may then argue as earlier in this paper that their variance is finite, that is, $Var(X_j^*) < \infty$ (j = 1, . . ., p; cf. Lord & Novick, 1968). Hence, the mean of each of these latent variables exists then, designated $\mathcal{E}(X_j^*)$, which we can denote T_j^* say and treat formally as a true score of X_j^*, just as in case X_j^* were to be itself observed (see preceding section; j = 1, . . ., p; cf. Kohli et al., 2014). With this in mind, for the jth underlying normal variable the following decomposition holds:

$$X_j^* = T_j^* + E_j^*, \tag{11}$$

where $E_j^* = X_j^* - T_j^*$ is set and could be formally considered an error score associated with X_j^* (j = 1, . . ., p).

An Alternative Congeneric Test Model for Binary Items

For a homogeneous multicomponent instrument consisting of binary or binary scored measures, X_1, . . ., X_p, one could argue in favor of the meaningfulness of the assumption that their underlying latent variables, X_1^*, . . ., X_p^*, are congeneric (see preceding discussion):

$$T_j^* = a_j^* + b_j^* T^*, \tag{12}$$

where T^* is their common underlying (true) score, while a_j^* and b_j^* are the associated intercepts and loadings (j = 1, . . ., p). We point out that Equation (12) represent a model that is testable using LVM, for instance, employing the weighted least squares approach for factor analysis with discrete variables (e.g., L. K. Muthén & Muthén, 2014). Its testability results from the fact that this model is not empirically distinguishable in the setting of interest in this article from the single-factor model with discrete indicators that itself is testable (e.g., Bartholomew, Knott, & Moustaki, 2011; see next for a qualification). We may also refer to the model defined in Equations (12) as a congeneric model with binary items, and to its special case with $b_1^* = \cdots = b_p^*$ as an essentially tau-equivalent model with binary items (see also Equations 13 and 14 below; cf. Kohli et al., 2014).

Since none of the underlying normal variables X_1^*, . . ., X_p^* is observed, however, their location parameters a_j^* are not uniquely estimable (identified) in the presence of their associated threshold parameters τ_j (j = 1, . . ., p). This underidentification issue is resolved by assuming a_j^* = 0 (j = 1, . . ., p; cf., e.g., Agresti, 2002). Based on this assumption, Equation (12) becomes

$$T_j^* = b_j^* T^* \quad (j = 1, \ldots, p) \tag{13}$$

With Equation (13), the definitional equations of the currently discussed congeneric model for binary items are

$$X_j^* = b_j^* T^* + E_j^*, \tag{14}$$

where one also assumes $Var(T^*)$ = 1 for identifiability reasons, and similarly that the error terms E_j^* are uncorrelated—that is, the covariance matrix of the error term vector $\underline{E}^* = (E^*_1, \ldots, E^*_p)'$ is diagonal—with its main diagonal elements denoted ψ_j (j = 1, . . ., p).

Classical Test Theory-Based Models Leading to Item Response Theory Models and Conversely

We denote next by $h(\underline{X}^*)$ the probability density function (pdf) of the random vector $\underline{X}^* = (X_1{}^*, \ldots, X_p{}^*)'$ of the underlying normal variables, and by $x_1, \ldots, x_p$ a series of 1s and 0s that represent a given response pattern on the original observed items $X_1, \ldots, X_p$ under consideration. (The following developments in this section that lead up to Equation (26) are a special case in the context of CTT of the more general FA-IRT relationship demonstrated in Takane & de Leeuw, 1987. They are presented here merely to emphasize a main point made in the Conclusion section, viz. that the choice between CTT and IRT in the setting of interest in this article results in actual fact from the choice of order of integration of the product of two appropriately defined functions; see Equation 17.)

Then because of the UNV assumption and definition of the underlying normal variables $X_1{}^*, \ldots, X_p{}^*$, the following relationship holds for the probability of observing that response pattern:

$$\mathrm{P}(X_1 = x_1, \ldots, X_p = x_p) = \int_I h(\underline{x}^*) d\underline{x}^*, \tag{15}$$

where I is an appropriate region of integration and $d\underline{x}^*$ is a shorthand for $dx_1^*, dx_2^* \ldots dx_p^*$.[1,2] Using the law or total probability (e.g., Raykov & Marcoulides, 2012), this pdf is on the other hand expressible as

$$h(\underline{x}^*) = \int_\Theta h(\underline{x}^* | T^*) g(T^*) dT^*, \tag{16}$$

where Θ denotes the space in which T^* varies, $h(\cdot|T^*)$ is the conditional pdf of $\underline{X}^*$ given T^*, and $g(\cdot)$ is the pdf of T^*.[3] With Equation (16), Equation (15) now becomes

$$\begin{aligned}\mathrm{P}(X_1 = x_1, \ldots, X_p = x_p) &= \int_I \left[\int_\Theta h(\underline{x}^* | T^*) g(T^*) dT^* \right] d\underline{x}^* \\ &= \int_I \int_\Theta h(\underline{x}^* | T^*) g(T^*) dT^* d\underline{x}^*. \end{aligned} \tag{17}$$

According to Fubini's integration theorem (e.g., Apostol, 2013), since the double integral in the right-hand side of Equation (17) exists, it is possible to change the order of integration in it. This leads to the following re-expression of the probability of the response pattern under consideration:

$$\mathrm{P}(X_1 = x_1, \ldots, X_p = x_p) = \int_\Theta g(T^*) \left[\int_I h(\underline{x}^* | T^*) d\underline{x}^* \right] dT^*. \tag{18}$$

Owing to Equation (14) and the uncorrelatedness of its error terms, the new inner integral in (18) can be rewritten as follows:

$$\int_I h(\underline{x}^*|T^*)d\underline{x}^* = \int_I \left[\prod_{j=1}^{p} h_j(x_j{}^*|T^*)\right] dx_1{}^*dx_2{}^* \cdots dx_p{}^*$$

$$= \prod_{j=1}^{p} \left[\int_{\tau_j}^{\infty} h_j(x_j{}^*|T^*)dx_j{}^*\right]^{x_j} \left[1 - \int_{\tau_j}^{\infty} h_j(x_j{}^*|T^*)dx_j{}^*\right]^{1-x_j}, \tag{19}$$

where $h_j(\cdot|T^*)$ denotes the conditional pdf of X_j^* given T^* ($j = 1, \ldots, p$).

However, because of Equations (10), (14), and the assumed normality of the underlying latent variable X_j^*,

$$\int_{\tau_j}^{\infty} h_j(x_j{}^*|T^*)dx_j = \Phi[(b_j{}^*T^* - \tau_j)/\psi_j] \tag{20}$$

holds ($j = 1, \ldots, p$). Therefore, from Equations (18) and (19) it now follows that the probability of the response pattern of consideration is

$$\mathrm{P}\left(X_1 = x_1, \ldots, X_p = x_p\right) =$$
$$\int_\Theta g(T^*) \prod_{j=1}^{p} \{\Phi[(b_j{}^*T^* - \tau_j)/\psi_j]\}^{x_j}\{1 - \Phi[(b_j{}^*T^* - \tau_j)/\psi_j]\}^{1-x_j} dT^*. \tag{21}$$

The right-hand side of Equation (21) is, however, precisely the expression for the probability of a given response pattern as represented by a two-parameter normal ogive item response model, up to notation used. Indeed, in that two-parameter normal ogive model,

$$\mathrm{P}\left(X_1 = x_1, \ldots, X_p = x_p\right) = \int_{-\infty}^{\infty} \mathrm{P}(X_1 = x_1, \ldots, X_p = x_p|\theta)u(\theta)d\theta \tag{22}$$

holds, owing also to the law of total probability, where $u(\theta)$ is the pdf of the random variable θ symbolizing the underlying ability evaluated by the used instrument with components X_1 through X_p. However, because of the local independence assumption typically advanced within the IRT framework,

$$\mathrm{P}(X_1 = x_1, \ldots, X_p = x_p|\theta) = \prod_{j=1}^{p} \mathrm{P}(X_j = x_j|\theta)$$

$$= \prod_{j=1}^{p} [p_j(\theta)]^{x_j}[1 - p_j(\theta)]^{1-x_j}, \tag{23}$$

where $p_j(\theta)$ denotes the item characteristic curve of the jth item, X_j ($j = 1, \ldots, p$). Next, in the two-parameter normal ogive model,

$$p_j(\theta) = \int_{-\infty}^{a_j\theta + b_j} \phi(z)dz = \Phi(a_j\theta + b_j) \tag{24}$$

holds, with $\phi(\cdot)$ being the pdf of the standard normal distribution, where a_j is the item discrimination parameter and b_j is the item difficulty parameter (using the typical IRT notation of discrimination and difficulty parameters; e.g., de Ayala, 2009). Hence, with Equations (23) and (24), Equation (22) states, in fact, that

$$P(X_1 = x_1, \ldots, X_p = x_p) = \int_{-\infty}^{\infty} u(\theta) \prod_{j=1}^{p} [\Phi(a_j\theta + b_j)]^{x_j} [1 - \Phi(a_j\theta + b_j)]^{1-x_j} d\theta. \tag{25}$$

As can be seen by direct comparison now, Equations (25) and (21) are identical with the following notational substitutions:

$$T^* = \theta,$$

$$b^*{}_j/\psi_j = a_j,$$

$$-\tau_j/\psi_j = b_j,$$

and

$$g(\cdot) = u(\cdot). \tag{26}$$

Hence, starting from the CTT-based congeneric test model (14) for binary items, via Equations (15) through (26) one obtains the associated probability for any prespecified response pattern on the measures X_1 through X_p, as *identical* to the probability for that pattern within the two-parameter normal ogive item response model. (Thereby, the second and third equations in (26) represent the relationship between their parameters; see also B. O. Muthén et al., 1991.) Conversely, by backtracking—that is, performing the above developments from (14) through (26) but in the reverse sequence—one obtains from the probability of any response pattern within the two-parameter normal ogive item response model, as *identical* that response pattern's probability associated with a corresponding CTT-based model, namely, a congeneric test model (14) with binary items.

The same argument of equivalence applies, along exactly the same lines, also for the one-parameter normal ogive model and the essentially tau-equivalent model with binary items, when the corresponding discrimination (loading) parameter equality is maintained throughout these developments for all items under consideration. Similarly, if the underlying latent variables X_j^* are assumed to begin with to be following a logistic distribution—practically equivalent to a corresponding normal distribution—then the same equivalence argument will be valid for the two- and one-parameter logistic item response models and corresponding CTT-based models (defined in the same way as the CTT-based models in the preceding section of this article). Thereby, if assuming a logistic distribution for each of the underlying variables to begin with, the only change in the preceding developments in this section will consist in substituting the cumulative distribution function $\Lambda(\cdot)$ of the standard logistic distribution wherever the cumulative distribution function $\Phi(\cdot)$ appears of the standard normal distribution.

This mutual implication of identity in the probability of any response pattern within CTT-based models on one hand and corresponding IRT-based models on the other hand (and vice versa), as shown above, represents the logical equivalence of the CTT and IRT frameworks that is in the center of interest of the present article.

Conclusion

For many years, CTT and IRT have represented the two major methodologies used for the purpose of test and scale construction and development in the educational, behavioral, and social sciences. The past several decades, however, have also witnessed substantial yet unjustifiable criticism of CTT and its potential for accomplishing this purpose. The concern of this article was with discussing in light of the extant methodological literature (Takane & de Leeuw, 1987) the equivalence between the CTT and IRT frameworks in the popular case of unidimensional multicomponent measuring instruments with binary or binary scored items and no guessing (see also Raykov & Marcoulides, 2011, ch. 11, 12). Unfortunately, because of the markedly technical nature of the topic, the main findings in Takane and de Leeuw (1987) have remained largely inaccessible and unclear for many empirical scientists. It was therefore the goal of this article to bring closer to the empirical behavioral and social scientists the CTT-IRT observational equivalence relation, which is obtainable from those prior findings, by employing the widely used setting of homogeneous dichotomous items. By doing so, following the far-reaching work by Takane and de Leeuw (1987) (see also Kamata & Bauer, 2008) on the more general FA-IRT relationship, we hoped to deal away with misconceptions and imprecise beliefs about the supposed deficiencies of CTT relative to IRT.

As demonstrated in the previous section, we wish to stress in this connection that the CTT-IRT (observational) equivalence for the setting of interest in this article results in the end from a *simple change in the order of integration* in a relevant double (two-dimensional) integral. This result, which was first demonstrated in a much more general setup relating FA and IRT in Takane and de Leeuw (1987), is itself rather revealing in our view and may well be seen as the very fundament of the CTT-IRT equivalence discussed in this article. In simple terms, it is in the end the *choice of the order of integration* (at times also referred to as ''marginalization'' process), which determines whether someone proceeds with an appropriate CTT-based model or a corresponding item response model, as observed in the preceding section.

We thus hope that with the present highlighting of the fundamental relationship between CTT and IRT and in particular its origin (see last two sections of this article), a main impediment for progress in the measurement field can be reduced significantly. This can be achieved in our opinion by disposing of the contemporary view concerning the assumed deficiency of CTT relative to IRT among some measurement researchers in the educational and behavioral disciplines. This view is unjustified and should be abandoned. Free from it and related misconceptions, researchers and students alike can embark on such a unified treatment, application, and use of CTT and IRT in the educational and behavioral disciplines, as well as beyond them, which combines their benefits rather than positions them against one another as in the recent and more distant past.

We conclude by pointing out that numerical illustrations of the CTT-IRT equivalence in the setting of concern to this article, whose theoretical justification was discussed and highlighted in the last two sections of the paper, can be found in the recent work by Kohli et al. (2014). Similarly, for a closely related discussion of the relationships between factor analysis and IRT in the more general case of multiple underlying latent variables evaluated by a set of ordinal measures, reference can be made to the instructive and generally applicable work by Takane and de Leeuw (1987; further related and insightful discussions in this respect are available, for instance, in Kamata & Bauer, 2008, and B. O. Muthén et al., 1991).

Acknowledgments

Thanks are due to R. Steyer, S. Penev, M. Edwards, C.-L. Lee, and C. Lewis for valuable discussions on classical test theory, factor analysis, and item response theory, as well as to an anonymous referee for critical comments on an earlier version of the article that have contributed considerably to its improvement.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The integration region I is the Cartesian product of the intervals (τ_j, ∞) for all measures with response 1, 'multiplied' with the Cartesian product of the intervals $(-\infty, \tau_j)$ for all measures with 0 response (across $j = 1, \ldots, p$). That is, if there are q positive and r zero responses in a pattern $(X_1, \ldots, X_p)$ under consideration $(q+r=p)$, then $\mathrm{I} = (\tau_1, \infty)$ x (τ_2, ∞) x … x (τ_q, ∞) x $(-\infty, \tau_{q+1})$ x $(-\infty, \tau_{q+2})$ x … x $(-\infty, \tau_r)$, where 'x' denotes Cartesian product, and without limitation of generality the first q responses in the pattern are taken to be 1's and the last r responses to be 0's.

2. As indicated earlier, this note is concerned with theoretical relationships between CTT and IRT rather than with numerical evaluation of integrals of relevance for it. In order to numerically approximate the integral in the last equation, and similarly the integrals following in the main text, one can make use for instance of Gauss-Hermite quadrature or related numerical methods (e.g., Stroud & Sechrest, 1966). Compared to analytically determined integrals, quadrature based integration can be seen as using a series of appropriately constructed 'parallelepipeds' to approximate respective multidimensional areas under the integrated (positive) functions - by summing up their 'volumes', numerical approximations are furnished in the end of the integrals of interest.
3. Since there are obviously no restrictions on T^*, which could result from the preceding discussion, the pertinent interval of integration is $\Theta = (-\infty, \infty)$.

References

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

Apostol, T. M. (2013). *Calculus*. New York, NY: Wiley.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis*. New York, NY: Wiley.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109-133.

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

Kohli, N., Koran, J., & Henn, L. (2014). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164414559071

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81-117.

Muthén, B. O., Kao, C.-F., & Burstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Applications of a new IRT-based detection technique. *Journal of Educational Measurement*, *28*, 1-22.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

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

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Raykov, T., & Marcoulides, G. A. (2012). *Basic statistics: An introduction with R*. New York, NY: Rowman & Littlefield.

Stroud, A. H., & Sechrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice Hall.

Takane, Y., & de Leeuw, J. (1987). On the relation between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Zimmerman, D. W. (1975). Probability measures, Hilbert spaces, and the axioms of classical test theory. *Psychometrika*, *40*, 221-232.