



Page Proof Instructions and Queries

Journal Title: APM
Article Number: 817779

Thank you for choosing to publish with us. This is your final opportunity to ensure your article will be accurate at publication. Please review your proof carefully and respond to the queries using the circled tools in the image below, which are available by clicking “Comment” from the right-side menu in Adobe Reader DC.*

Please use *only* the tools circled in the image, as edits via other tools/methods can be lost during file conversion. For comments, questions, or formatting requests, please use . Please do *not* use comment bubbles/sticky notes .



*If you do not see these tools, please ensure you have opened this file with **Adobe Reader DC**, available for free at get.adobe.com/reader or by going to Help >Check for Updates within other versions of Reader. For more detailed instructions, please see us.sagepub.com/ReaderXProofs.

Sl. No.	Query
	GQ: Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.
	GQ: Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
	GQ: Please confirm you have reviewed this proof to your satisfaction and understand this is your final opportunity for review prior to publication.
	GQ: Please confirm that the Funding and Conflict of Interest statements are accurate.
1	Please check whether the affiliation and the corresponding author details are correct.
2	As per APM style, first person pronouns “we,” “our,” and “us” are not allowed in the text, but the terms “we” and “our” stated throughout the article seem to indicate the reader as well as the author, hence they have been retained as given. But in certain places, “we” has been changed to either to passive constructions or replaced with “the authors.” Please indicate if it refers only to the authors in any of the context. Also check whether the edits made to the sentences retain the intended meaning, where first persons are used throughout the article.
3	Please check whether the expansion given for “TM” is correct.
4	Note that “c.d.f.” has been changed to “CDF.” Also please check whether the inserted expansion is correct.
5	Please provide expansion for “DTM.”
6	“Conijn et al., 2016” is not listed in the references. Please provide reference details or delete the citation.
7	Note that the citation “Yuan et al., 2017” has been changed to “Yuan, Chan, Marcoulides, & Bentler, 2016” as per reference list. Please check and confirm.
8	Please check whether the formatting and edits made in Tables 1, 2, A1, and A2 are correct.
9	Please provide equation after “For the DTGRM” in the appendix.

- 10 “Muthén & Kaplan, 1985” is not listed in the references. Please provide reference details.
 - 11 Please provide expansion for “D.D.” in Table A1, if appropriate.
 - 12 “Mislevy, 1984” is not cited in text. Please indicate where a citation should appear or allow us to delete the reference.
 - 13 Please check whether the edits made to the reference “Pallero et al., 1998” are correct.
-



A Comprehensive IRT Approach for Modeling Binary, Graded, and Continuous Responses With Error in Persons and Items

Pere J. Ferrando¹

Abstract

Dual item response theory (IRT) models in which items and individuals have different amounts of measurement error have been proposed in the literature. Any developments in these models, however, are feasible only for continuous responses. This article discusses a comprehensive dual modeling approach, based on underlying latent response variables, from which specific models for continuous, graded, and binary responses are obtained. Procedures for (a) calibrating the items, (b) scoring individuals, (c) assessing model appropriateness, and (d) assessing measurement precision are discussed for all the resulting models. Simulation results suggest that the proposal is quite feasible. A practical illustration is given with an empirical example in the personality domain.

Keywords

personality measurement, person reliability, item discrimination, factor analysis, item response theory

In the psychometric models commonly used in typical-response (personality and attitude) measurement, such as linear factor analysis (FA), the graded-response model (GRM), and the two-parameter model (2PM), items are characterized by two types of parameter: location and discrimination. Individuals, however, are only characterized by one location parameter (position on the trait continuum). Theory and evidence, however, suggests that this modeling is insufficient (Tellegen, 1988). Just as items generally differ in their sensitivity at differentiating between individuals with different trait levels, individuals also generally differ in the sensitivity of their responses to the different item locations. Some respondents are largely insensitive, and their response patterns are almost random. At the opposite extreme, some individuals respond with high consistency, leading to response patterns that approach Guttman patterns (Ferrando, 2004, 2013; Fiske, 1968). If this scenario is accepted, then a “dual” modeling (see Fiske, 1968) in which both items and persons differ in terms of discriminating power seems to be the most plausible approach to fitting typical responses. [AQ: 2]

¹Universitat Rovira i Virgili, Tarragona, Spain

Corresponding Author:

Pere Joan Ferrando, Facultat de Psicologia, Universitat Rovira i Virgili, Carretera Valls s/n, Tarragona, 43007,

Spain. [AQ: 1]

Email: perejoan.ferrando@urv.cat

Dual models of the type discussed above have been discussed in the literature since the 1940s (Mosier, 1942), although the purposes of these discussions and the terminology used are often quite different (see, for example, Ferrando, 2004). A review, however, suggests that these models can be divided into two main families. The models in the first family are Thurstonian (TMs), which model individual discrimination (or individual error) as random fluctuation around a central trait level (Ferrando, 2004, 2007, 2009; Levine & Rubin, 1979; Lumsden, 1980) [AQ: 3]. Models in the second family are multiplicative models (MMs), which model individual discrimination as a person slope that functions multiplicatively with the item slope (Ferrando, 2014, 2016; Lubbe & Schuster, 2016, 2017; Strandmark & Linn, 1987).

Both TMs and MMs were initially considered only for binary responses, and, in this format, both lead to very similar outcomes and interpretations. Extension to more continuous formats, however, is more complex. Although the person discrimination parameter in TMs has the same interpretation in any format (as discussed below), the person slope in the MMs can also be thought to model individual differences in response scale usage (Ferrando, 2014) or proneness to extreme responding (Lubbe & Schuster, 2016, 2017) in the case of continuous or graded formats. In this sense, MMs are less specific and more difficult to interpret than TMs (e.g., van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011).

From an applied point of view, there are feasible procedures for fitting dual MMs for binary (Ferrando, 2016), graded (Lubbe & Schuster, 2017), and continuous (Ferrando, 2014; Lubbe & Schuster, 2016) responses. However, to date this is not the case for dual TMs, for which a full feasible procedure has only been proposed for continuous responses (Ferrando, 2013). Dual TMs for binary and graded responses have been considered intractable in practice (Lumsden, 1980; Torgerson, 1958), and only restricted versions in which item discrimination is constant appear to exist at present (Ferrando, 2004, 2007, 2009).

The main purpose of this article is to propose a comprehensive, item response theory (IRT)-based, dual TM approach that can be used with binary, graded, and continuous typical-response items. The resulting specific models are denoted as DTCRM (continuous responses; the already existing model), DTGRM (graded responses), and DTBRM (binary responses). For the DTGRM and DTBRM, the practical limitations mentioned above are overcome by using an underlying variables approach (UVA, Muthén, 1984), which makes the processes of fitting and scoring these models quite feasible in practice. So, the present proposal is mainly applied, and practical procedures are proposed for (a) calibrating the items and assessing model-data fit and appropriateness, (b) estimating the person parameters (scoring), and (c) assessing the precision with which the individual parameters are estimated. To the best of the author's knowledge, the UVA-based developments that lead to the DTGRM and DTBRM are new contributions, and so are the specific procedures that are proposed for calibrating the items and scoring the individuals (although they are indeed specific applications of more general, well-known procedures). Finally, the analytical results concerning the precision of the individual parameter estimates also seem to be new.

The DTCRM: A Review

Consider a test made up of $j = 1 \dots n$ items with an approximately continuous format that is administered to a sample of $i = 1 \dots N$ individuals. The test aims to measure a trait θ , assumed to have zero mean and unit variance in the population, and, for interpretative purposes, the item scores are scaled to have values between 0 and 1. Let X_{ij} be the score of individual i in item j . The structural model for this score is

$$X_{ij} = 0.5 + \lambda_j(T_i - b_j), \quad (1)$$

where T_i is the momentary trait (or perceived trait) value of respondent i when answering item j , and b_j is the momentary (perceived) location of item j on the trait continuum.

$$T_i = \theta_i + \omega_i; \quad b_j = \beta_j + \varepsilon_j. \quad (2)$$

For a given respondent i , consider first the distribution of T_i over the test items. This distribution is assumed to be normal with mean θ_i and variance σ_i^2 , which are the parameters that characterize respondent i , and that remain constant across items. Now, θ_i and σ_i^2 generally take on different values for different individuals, and they are assumed to be independent random variables over respondents. As for item j , the distribution of b_j , over respondents is assumed to be normal, with mean β_j , and variance $\sigma_{\varepsilon_j}^2$. Finally, the item and person residuals are assumed to be independent (e.g., Torgerson, 1958). Regarding terminology and interpretation, θ_i , denoted here as person location, is Mosier's (1942) "respondent characteristic value," the single value that best summarizes the standing of individual i on the trait. For its part, β_j , denoted here as item location, can be interpreted as a conventional IRT difficulty index, as discussed below. As for the error variance terms, Thurstonian terminology is used, and σ_i^2 is referred to as the person discriminial dispersion (PDD) and σ_{ε}^2 as the item discriminial dispersion (IDD). The PDD is a direct measure of trait variability. As for the IDD, it is usually related to the degree of item ambiguity, but it might also depend on general characteristics such as type of stem and average length (e.g., DeFleur & Catton, 1957; Ferrando, 2013; Lumsden, 1980; Taylor, 1977).

From the conditions discussed so far, it follows that the conditional distribution of X_j for fixed θ_i and σ_i^2 is normal, with expectation and variance given by

$$E(X_{ij}|\theta_i, \sigma_i^2) = 0.5 + \lambda_j(\theta_i - \beta_j); \quad \text{Var}(X_{ij}|\theta_i, \sigma_i^2) = \lambda_j^2(\sigma_i^2 + \sigma_{\varepsilon_j}^2). \quad (3)$$

The expressions in Equation 3 can be interpreted, respectively, as the expected mean and variance of X_j across all respondents with person location θ_i and PDD σ_i^2 . An alternative interpretation is to view them as the expected mean and variance of the scores of respondent i across items with the same parameters as item j .

The conditional expected score in Equation 3 is a direct function of the weighted person-item distance $\lambda_j(\theta_i - \beta_j)$. When $\theta_i > \beta_j$, the expected score is above the 0.5 response scale midpoint (i.e., 0.5), and when the person location matches the item location, the expected item score is the midpoint. So, as proposed above, β_j can be interpreted as a standard IRT difficulty index: It is the point on the trait continuum that marks the transition from the tendency to disagree/not endorse the item to the tendency to agree/endorse it.

At this point, it might be of interest to compare the expectations in Equation 3 to the expectation derived from the linear MM for continuous responses (Ferrando, 2014; Lubbe & Schuster, 2016).

$$E(X_j|\theta_i, \omega_i) = 0.5 + \xi_i \lambda_j(\theta_i - \beta_j). \quad (4)$$

As in the DTCRM, the expected response in the MM Equation 4 is also a direct function of the person-item distance. However, the person parameter ξ_i in Equation 4 (assumed to be positive) acts as a moderator that amplifies or reduces the impact of this distance on the expected item response. So, for large ξ_i , a small positive distance leads to an expected response that goes toward the upper end of the item response scale. This functioning was initially intended to model individual discrimination (in terms of sensitivity to the person-item distance). However, it might also well reflect idiosyncratic responding (proneness to extreme responding) or lack of cognitive effort. In contrast, in the DTCRM modeling, the PDD does not affect the extremeness

of the expected response but does affect its consistency. Thus, when both PDD and IDD are small, so is the conditional variance in Equation 3, which means that the observed score is close to the expected score.

By recalling now that the marginal mean and variance of θ are assumed to be 0 and 1, respectively, it follows that the marginal mean and variance of X_j over the entire population of respondents are

$$E(X_{ij}) = 0.5 - \lambda_j \beta_j = \mu_j; \quad \text{Var}(X_{ij}) = \lambda_j^2 [\text{Var}(\theta) + E(\sigma_i^2) + \sigma_{\varepsilon j}^2] = \lambda_j^2 [1 + E(\sigma_i^2) + \sigma_{\varepsilon j}^2]. \quad (5)$$

And the product-moment correlation between the scores on items j and k is

$$\rho(X_j, X_k) = \alpha_j \alpha_k, \quad (6)$$

where

$$\alpha_j = \frac{\lambda_j}{\sqrt{\text{Var}(X_j)}} = \frac{1}{\sqrt{1 + E(\sigma_i^2) + \sigma_{\varepsilon j}^2}} \quad (7)$$

is the correlation between the scores on item j and θ (standardized loading in FA terminology).

The simple linear model reviewed in this section assumes that X_{ij} is bounded whereas T_i and b_j are not. So, the model cannot be strictly correct and must necessarily be considered as an approximation. More in detail, the assumptions above imply that (a) the item response function is nonlinear rather than linear, and (b) the conditional distributions become more asymmetrical and with decreased variance toward the ends of the scale. In most practical applications, however, especially in personality measurement, the linear model as an approximation is expected to work reasonably well (see Ferrando, 2002, for a discussion).

The DTGRM and DTBRM

Consider now that the observed item score X_j is a categorical variable, and assume that (a) there is a latent response variable Y_j that underlies X_j , and (b) the following model holds for Y_j

$$Y_{ij} = \alpha_j (T_i - b_j), \quad (8)$$

where T_i and b_j behave as in Equation 2. Equation 8 is the same model as Equation 1 without the midpoint intercept term and with the variance of Y_j fixed to 1, which means that the scale parameter λ_j is now a standardized loading α_j (see Equation 7). This variance restriction is because, in contrast to Equation 1, the origin and scale for Y_j are now undetermined. In the standard UVA modeling (e.g., Muthén, 1984), this indeterminacy is usually solved by assuming that the marginal distribution of Y_j is normal with zero mean and unit variance. In the present modeling, the unit variance assumption has already been adopted. As for the normality assumption, the marginal distribution of Equation 8 for a fixed item is that of the sum of three independent variables (θ , ε , and ω ; see Equation 2), of which θ and ε are normal, and ω follows a Pearson type-VII distribution (see Ferrando, 2007). For practical purposes, the resulting distribution is close enough to normal for this assumption also to be used. The mean of Y_j , however, cannot be assumed to be generally zero. In effect, the marginal mean and variance are given by

$$E(Y_j) = -\alpha_j \beta_j = \mu_j; \quad \text{Var}(Y_j) = 1 = \alpha_j^2 [1 + E(\sigma_i^2) + \sigma_{\varepsilon j}^2]. \quad (9)$$

And the product-moment correlation between the latent scores on items j and k is

$$\rho(Y_j, Y_k) = \alpha_j \alpha_k, \quad (10)$$

where α_j is the product-moment correlation between the latent scores on item j and the central trait level θ .

The relation between Y_j and the observed score X_j is now assumed to be a step function governed by a threshold mechanism. The most usual scoring schemas for categorical variables are considered here: 0 and 1 for the binary case, and integer values 1, 2, . . . for the graded-response case. With this schema, the mechanism is

$$\begin{aligned} X &= 0 & \text{if } Y < \tau \\ X &= 1 & \text{otherwise} \end{aligned} \quad (11)$$

for the binary case, and

$$\begin{aligned} X &= 1 & \text{if } Y < \tau_1 \\ X &= 2 & \text{if } \tau_1 \leq Y < \tau_2 \\ X &= 3 & \text{if } \tau_2 \leq Y < \tau_3 \\ & \vdots \\ X &= c & \text{if } \tau_{c-1} < Y \end{aligned} \quad (12)$$

for the graded-response case with c response categories. From this modeling, it follows that the product-moment correlation between Y_j and Y_k is the tetrachoric correlation between X_j and X_k in the binary case, and the polychoric correlation between X_j and X_k in the graded-response case.

We turn now to the IRT modeling implied by the UVA described so far. In the DTBRM, the probability of endorsing item j for fixed θ_i and σ_i^2 is

$$P(X_{ij} = 1 | \theta_i, \sigma_i^2) = \Phi \left(\frac{1}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \left(\theta_i - \left(\beta_j + \frac{\tau_j}{\alpha_j} \right) \right) \right) = \Phi(\gamma_{ij}(\theta_i - \delta_j)), \quad (13)$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution [AQ: 4]. To see the role of the person parameters in Equation 13, note that θ_i determines which score (0 or 1) is the most probable for this respondent. As for the PDD, when σ_i^2 decreases, the responding becomes more deterministic and sensitive to the item location: The respondent tends to endorse the item if its location is below θ_i (“easy” item for him or her) and reject the item if it is above θ_i (“difficult” item).

In the DTGRM, the probability of scoring k in item j for fixed θ_i and σ_i^2 is

$$\begin{aligned} P(X_{ij} = k | \theta_i, \sigma_i^2) &= \Phi \left(\frac{1}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \left(\theta_i - \left(\beta_j + \frac{\tau_{jk-1}}{\alpha_j} \right) \right) \right) - \Phi \left(\frac{1}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \left(\theta_i - \left(\beta_j + \frac{\tau_{jk}}{\alpha_j} \right) \right) \right) \\ &= \Phi(\gamma_{ij}(\theta_i - \delta_{jk-1})) - \Phi(\gamma_{ij}(\theta_i - \delta_{jk})). \end{aligned} \quad (14)$$

In Equation 14, the person location θ_i determines the response category that has the greatest probability of being endorsed by respondent i . As for the role of PDD, consider a respondent whose person location is between δ_{jk-1} and δ_{jk} . As the PDD approaches zero, the probability of endorsing category k increases, whereas the probability of endorsing the remaining categories decreases. So, the process of responding becomes more deterministic. At the opposite extreme, as the PDD increases, the probability of responding in different categories becomes progressively more undifferentiated. This way of working contrasts with that of MM for graded responses (Lubbe & Schuster, 2017). As in the linear case, the person slope parameter in the MM modifies the expected response, so that low slope values imply that the response is more likely to lie in the middle categories whereas with large values it is more likely to be in the outer categories (see Lubbe & Schuster, 2017, for details). Again, this might reflect either person discrimination or idiosyncratic responding.

In the literature, the DTBRM in Equation 13 is Lumsden's (1980) "Two-parameter 3 model," which he considered to be the most general model intended for binary items. If the PDDs are equal for all respondents (i.e., $\sigma_i^2 = \sigma^2$) while the IDDs are allowed to vary, Equation 13 becomes equivalent to the standard 2P normal-ogive model. And, if the IDDs are equal for all the items but the PDDs are different for different individuals, Equation 13 reduces to Lumsden's (1980) "Pseudo-Rasch 2 model" or Torgerson's (1958) Condition C, which is the model considered by Ferrando (2004, 2007). In the graded-response case, the DTGRM in Equation 14 reduces to the normal-ogive version of Samejima's (1969) GRM under the first restriction, and to Ferrando's (2009) graded model under the second. It is noted finally that all the normal-ogive models discussed so far are obtained by using a formulation other than the standard one in IRT (e.g., Lord & Novick, 1968). Torgerson (1958) called this formulation the "alternative general normal-ogive model."

Fitting the DTMs[AQ: 5]

The general approach proposed for all the models considered is a conventional two-stage conditioned procedure (McDonald, 1982) with a first calibration stage in which the item parameters are estimated, and a second scoring stage in which estimates of the person locations and the PDDs are obtained for all the individuals. In addition, a multifaceted approach is proposed for assessing the appropriateness of the fitted model.

Item Calibration

The three models can be fitted by using a limited-information FA approach with additional identification restrictions. A unified approach is proposed in which items are calibrated by fitting the unidimensional FA model to the appropriate inter-item correlation matrix: Product-moment (DTCRM), tetrachoric (DTBRM), and polychoric (DTGRM). The basic approach is standard, so specific estimation procedures and discrepancy functions will not be discussed here, although some discussion is provided in the example.

The main estimates obtained by fitting the FA model are the standardized loadings α and the corresponding standardized residual variances. Now, for all three models, the following result is obtained (see Equations 7 and 9).

$$\frac{1 - \alpha_j^2}{\alpha_j^2} = E(\sigma_i^2) + \sigma_{\varepsilon j}^2. \quad (15)$$

Equation 15 means that the inter-item correlation matrix does not contain sufficient information to separately identify the average PDD and the IDD. In the dual MMs, this problem is settled by fixing the mean person slope to 1 (Ferrando, 2014). This constraint, however, cannot be used here, because all the IDDs must be greater than 0. So, the identification approach proposed in this case is based on the use of a marker variable. The “best” item (i.e., the item with the largest standardized loading) is chosen as a marker and so treated as if its IDD was zero. Then, relative to this scaling, the average PDD is estimated as

$$\frac{1 - \hat{\alpha}_{(\max)}^2}{\hat{\alpha}_{\max}^2} = \hat{E}(\sigma_i^2), \quad (16)$$

where $\hat{\alpha}_{(\max)}$ is the largest estimated standardized loading. The remaining IDDs are obtained from Equation 15.

We turn now to the item location parameters. In the case of continuous scores, they can be estimated directly from the marginal means (see Equation 5). In the case of binary scores, conventional fitting of the 2PM using the UVA approach will provide estimates of the transformed location parameters δ_j in Equation 13. However, β_j cannot be identified separately from δ_j because the origin of Y_j is undetermined. Now, in principle, β_j does not need to be identified separately to obtain individual PDD estimates at the scoring stage. However, it can be by assuming that items are categorized at a common threshold of 0 in Equation 11. This fixes the origin of Y_j and provides a plausible interpretation (see Lubbe & Schuster, 2017): negative values of Y_j lead to denial while positive values lead to item endorsement.

In the graded-response case, the β_j s can be identified by extending the above rationale in the way proposed by Lubbe and Schuster (2017). If the number of categories is even, the middle threshold is fixed to 0 for all the items. If it is odd, the sum of the two central thresholds is fixed to 0. Again, β_j does not need to be identified to obtain PDD estimates in the DTGRM. However, identification is useful for interpretative purposes because, as occurs with the DTBRM and the DTCRM, it also provides a single item location in the graded response case.

Scoring

In the original linear model, Ferrando (2013) proposed using maximum likelihood (ML) to estimate the person parameters. Experience suggests that ML estimation is not only feasible but also prone to giving some very large PDD estimates if the test is short or the item locations are not evenly distributed. This problem becomes worse in the DTGRM, and more so in the DTBRM.

To overcome the problem above, the approach proposed here is to use Bayes expected a posteriori (EAP, Bock & Mislevy, 1982) estimation for all the models considered. This procedure has two main advantages. First, it uses the mean PDD estimate obtained in the calibration stage to center the prior distribution. So, the “ensemble biases” (Mislevy, 1986) phenomenon of shrinkage toward an inappropriate central value is avoided. Second, it ensures that the person estimates (especially σ_i^2) fall within reasonable values. EAP estimation of θ and σ_i^2 in all the models is conventional and is detailed in the appendix.

From a modeling point of view, the most important issue in the EAP estimation process is the choice of the prior distributions. In our proposal, the prior for θ is set as standard normal by default, but estimated distributions via quadrature can also be used as input (Mislevy, 1986). As for the PDDs, they are variances, so their most appropriate prior is the scaled inverse χ^2 distribution (Novick & Jackson, 1974). Because only the prior mean is estimated at the calibration stage, the prior variance for σ_i^2 is indeterminate. So, a prior distribution for σ_i^2 needs to be

specified so that plausible estimates can be obtained for all the respondents but, at the same time, it should not be so tight that it produces excessive regression toward the prior mean. This point is also discussed in the appendix.

For each individual, the output of the EAP procedure consists of the θ_i and σ_i^2 point estimates and the corresponding posterior standard deviations (PSDs) which serve as standard errors (e.g., Bock & Mislevy, 1982). For both θ_i and σ_i^2 , a conditional PSD-based reliability estimate can further be obtained as

$$\begin{aligned}\rho(\hat{\theta}_i) &= 1 - \frac{\text{PSD}(\hat{\theta}_i)^2}{\text{Var}(\theta)}, \\ \rho(\hat{\sigma}_i^2) &= 1 - \frac{\text{PSD}(\hat{\sigma}_i^2)^2}{\text{Var}(\sigma^2)}.\end{aligned}\tag{17}$$

Finally, an empirical marginal reliability estimate can be obtained by averaging the squared PSDs in the sample of N individuals (Brown & Croudace, 2015):

$$\begin{aligned}\rho(\hat{\theta}) &= 1 - \frac{\sum_i^N [\text{PSD}(\hat{\theta}_i)]^2}{N\text{Var}(\theta)}, \\ \rho(\hat{\sigma}^2) &= 1 - \frac{\sum_i^N [\text{PSD}(\hat{\sigma}_i^2)]^2}{N\text{Var}(\sigma^2)}.\end{aligned}\tag{18}$$

Provided that the PSDs remain relatively uniform, the marginal reliabilities in Equation 18 are representative of the overall precision of the estimates in the population of respondents.

Assessing Model Appropriateness

In all the models proposed here, calibration consists of fitting a unidimensional FA model. So, model-data fit and appropriateness at this level can be assessed by using standard procedures. Appropriate fit of the FA model, however, is necessary but not sufficient, because the DTMs cannot be distinguished from the corresponding normative (i.e., constant PDD) models in terms of their implied inter-item correlation matrices. So, further procedures are needed to decide whether the more flexible but also more parameterized dual TM provides a non-negligibly better fit than the corresponding model with constant PDD.

The approach proposed has been systematically used in previous related models, and is based on a likelihood ratio (LR) statistic. For a single respondent i , let $L_i^0(\hat{\theta}_i, \hat{\sigma}^2)$ be the value of the likelihood function evaluated by using the person location estimate that is obtained under the restriction that all the PDDs have a constant value. Now, let $L_i^1(\hat{\theta}_i, \hat{\sigma}_i^2)$ be the corresponding value using both the person location and the PDD estimate (see the appendix for further details). The LR statistic and the transformed value proposed are

$$\Lambda_i = \frac{L_i^0(\hat{\theta}_i, \hat{\sigma}^2)}{L_i^1(\hat{\theta}_i, \hat{\sigma}_i^2)}; s_i = -2 \ln(\Lambda_i).\tag{19}$$

Statistic Λ_i is a descriptive normed index with values in the range 0 to 1. Values close to 0 indicate that the dual TM provides a substantially better fit than the corresponding standard model. As for s_i , under very restrictive conditions, it could be considered to be a value randomly drawn

from a χ^2 distribution with one degree of freedom. And, by further assuming experimental independence between respondents, the sum $Q = \sum s_i$ asymptotically approaches a χ^2 distribution with N degrees of freedom (see Ferrando, 2013). However, for this being so, the likelihoods must be evaluated at their ML estimates whereas here they are evaluated at their EAP estimates. To sum up, Q has been proposed as the overall index for assessing whether the DTM fits better than its standard counterpart but acknowledge that it cannot be used as a strict inferential measure and that the reference distribution is at best only an approximate guide. The behavior of the statistic is assessed below via simulation.

Substantive and Practical Considerations

The DTMs are not only more flexible than their normative counterparts but also more complex and potentially prone to producing unstable parameter estimates. Therefore, the conditions in which the proposed models are appropriate and expected to work well in practice need to be discussed.

Analytical expressions for the θ_i and σ_i^2 PSDs in the three models are provided in the appendix (Equations 30 to 32). For both parameters, accuracy of the estimates increases with test length and (in the DTGRM) with the number of categories (see Ferrando, 2009). However, accurate estimation of θ_i requires items whose locations are close to the parameter value, whereas accurate estimation of σ_i^2 requires items with locations that are far removed from θ_i . So, the “ideal” scenario is a long test with item locations that are widely spread and evenly distributed around the mean person location value (zero in the present scaling). These are also the “ideal” conditions of any broad-bandwidth personality test. Extensive simulation is needed before any recommendations are given, but for the moment, it is important to inspect the PSDs and reliabilities of the individual estimates to check that they are accurate for most of the respondents.

In comparative terms, the θ_i estimates are expected to be substantially more reliable than the σ_i^2 estimates. This result has been systematically obtained in the literature (Ferrando, 2004; Mosier, 1942), and, in the present proposal, can be obtained by using Equations 30 to 32 in the appendix. The situation is the same that occurs when estimating item slopes in the 2PM with the role of the items and respondents reversed: In general, the location estimate is far more reliable than the slope estimate (e.g., Lord & Novick, 1968). Even so, however, the results obtained in the appendix for the σ_i^2 estimates suggest that, in a well-designed test with a good spread of item locations and medium to high item discriminating power, reliabilities of 0.70 can be reached with about 25 items, and of 0.80 with about 40 items.

We turn now to the potential advantages of using the DTMs. To start with, they provide additional information about the consistency of the respondent’s answering behavior via the PDD estimate. This information, in turn, can be of use in individual assessment or in exploratory person-fit research (see Conijn et al., 2016). Furthermore, it has been hypothesized that the PDD is related to the relevance and degree of clarity and strength with which the trait is internally organized in the individual (Traitedness; for example, Markus, 1977; Reise & Waller, 1993; Taylor, 1977; Tellegen, 1988). Evidence based on previous TM-based applications or related indices suggests that the PDD estimates can be effectively used to reflect traitedness (LaHuis, Barnes, Hakoyama, Blackmore, & Hartman, 2017; Reise & Waller, 1993). [AQ: 6]

As for the role of PDDs in individual assessment, the accuracy with which the person locations are estimated is better assessed with the DTMs (provided they are correct). Indeed, the analytical expressions for the θ_i PSDs provided in the appendix are

$$\frac{1}{\text{PSD}^2(\hat{\theta}_i)} \cong 1 + \sum_j^n \frac{1}{\sigma_i^2 + \sigma_{ej}^2} (\text{DTCRM}), \quad (20)$$

$$\frac{1}{\text{PSD}^2(\hat{\theta}_i)} \cong 1 + \sum_j^n \left(\frac{1}{\sigma_i^2 + \sigma_{ej}^2} \right) \frac{\phi^2 \left[\left(\frac{(\hat{\theta}_i - \delta_j)}{\sqrt{\sigma_i^2 + \sigma_{ej}^2}} \right) \right]}{P_{ij} Q_{ij}} (\text{DTBRM}),$$

$$\frac{1}{\text{PSD}^2(\hat{\theta}_i)} \cong 1 + \sum_j^n \left(\frac{1}{\sigma_i^2 + \sigma_{ej}^2} \right) \sum_k^c \frac{1}{P_{ijk}} \left[\left[\phi \left(\frac{(\hat{\theta}_i - \delta_{j,k-1})}{\sqrt{\sigma_i^2 + \sigma_{ej}^2}} \right) - \phi \left(\frac{(\hat{\theta}_i - \delta_{j,k})}{\sqrt{\sigma_i^2 + \sigma_{ej}^2}} \right) \right]^2 \right] (\text{DTGRM}),$$

where ϕ is the density of the standard normal distribution, $P_{ijk} = P(X_{jk}|\theta_i, \sigma_i^2)$, and $Q = 1 - P$. In the three models, the accuracy of the person location θ estimate depends on the amount of PDD. All other things being constant, the θ estimates are more accurate for the most discriminating and reliable individuals (see Equation 17). This result is particularly important when θ is estimated in clinical settings or in selection or classification processes.

Finally, for psychometric and conceptual reasons, the PDD are expected to have a moderating role in validity assessment. First, as discussed above, the person location estimates of the less discriminating individuals are less reliable and, from basic attenuation theory, the unreliability of the score estimates attenuates the validity coefficient (e.g., Lord & Novick, 1968). Second, those individuals for whom the trait is relevant are expected to be more likely to display a stronger correspondence between trait self-description and external trait-relevant variables (Markus, 1977; Paunonen, 1988). For both reasons, those individuals with smaller PDDs would tend to be the most predictable although, in practice, the differential validity effects are expected to be modest at best (Ferrando, 2004, 2013).

The relatively low degree of reliability of the σ_i^2 estimates is admittedly a limiting factor for their potential usefulness, especially if these estimates were to be used for accurate individual assessment of the person discrimination levels. For the auxiliary roles discussed above or for validity assessment, however, previous results (Ferrando, 2004, 2009, 2013) suggest that if minimally acceptable reliabilities of about 0.70 can be obtained (which is reasonable in a good designed study), then the σ_i^2 estimates discrimination is already appropriate.

Simulation Studies

Experience with the DTCRM suggests that the limited-information procedure proposed in this article works quite well in the simple linear case. For this reason, two simulation studies that focused on the new approaches proposed here as well as on the main differences with the original DTCRM proposal were undertaken. More specifically, the study considered only the DTGRM, which is the most general of the two UVA-based models. Due to space limitations, the complete studies as well as the tables of results (Tables A1(a), A1(b), and A2) are presented in the appendix, and only a summary is provided here.

The first study assessed the extent to which the approaches proposed here provide appropriate item calibration results and, above all, acceptable individual estimates of the two types of person parameter. Thus, results are presented at two levels: calibration and scoring. In the first calibration stage, the main aim was to check that data generated by the DTGRM did in fact behave like an FA model at the correlational level and that items could be well calibrated by fitting Spearman's model to the inter-item polychoric correlation matrix.

The scoring part of the study is of more interest because now Bayes EAP estimates are used instead of the ML estimates originally proposed. The focus here was on the appropriate recovery of the “true” individual parameters and on the accuracy of the individual estimates.

The calibration results suggested that the FA model provided a good fit in all cases and appropriate recovery of the item parameters. The scoring results were also positive. For both θ_i and σ_i^2 the parameters were well recovered within the accuracy limits discussed in the section above. Furthermore, the empirical and model-based accuracy results agreed reasonably well.

The second study aimed to assess the behavior of the LR Q statistic proposed above when the likelihoods are evaluated at their EAP estimates. Two situations were considered: H_0 , in which the correct model was the standard GRM with constant PDD, and H_1 in which the correct model was the DTGRM. The results suggested that (a) the statistic allowed the correct model to be distinguished in all conditions and (b) power increased with test length, as expected. However, the statistic was conservative, and, under H_0 it provided values systematically smaller than the chi-square expectations. This point is discussed further below.

Illustrative Example

Ferrando (2013) illustrated the functioning of the DTCRM with an instrument known as CTAC, a Spanish acronym for “Anxiety Questionnaire for Blind People.” The CTAC (Pallero, Ferrando, & Lorenzo-Seva, 1998) is a 35-item test that measures anxiety in situations related to visual deficit and which is intended to be used in the general adult population with severe visual impairment. The response format is 5-point Likert-type and, in the population for which the test is intended, the distributions of the item scores are generally unimodal and not extreme. This result suggests that, “a priori,” both the DTCRM and the DTGRM may be appropriate (see Culpepper, 2013). So, the results they provide can be compared for illustrative purposes. In Ferrando’s (2013) example, the CTAC was fitted in a sample of 352 respondents. Here a far larger sample of 760 adults collected from various centers belonging to the Spanish National Organization of the Blind (ONCE) is used.

The unidimensional FA model was fitted to the product-moment (DTCRM) and polychoric (DTGRM) inter-item correlation matrices by using robust unweighted least squares (ULS) estimation as implemented in the FACTOR program (Lorenzo-Seva & Ferrando, 2013). Appropriateness and goodness of fit were assessed at this stage by using a multifaceted approach that includes (a) conventional goodness-of-fit assessment, (b) equivalence testing as proposed by Yuan, Chan, Marcoulides, and Bentler (2016) (only available at present for the continuous model), and (c) measures of strength and replicability of the solution as well as closeness to unidimensionality. For both models, the results are in Table 1. They are clear and can be summarized as follows: The fit is quite acceptable by all standards, the solutions are strong and replicable, and the data are essentially unidimensional. As expected, the results for the continuous and the graded models at this stage are very similar [AQ: 7]

The results for the LR test are at the bottom of Table 1. Even with the limitations of the Q statistic discussed above, they are quite clear, and more so given the conservative behavior of the test. In both the continuous and graded case, they suggest that the DTM is more appropriate than the corresponding normative model.

The calibration results for both models are now summarized. The standardized weights (α) ranged from 0.58 to 0.70 (DTCRM) and 0.62 to 0.74 (DTGRM), which are quite acceptable for personality items. The product-moment correlation between the α estimates produced by both models was .99. So, as expected, (a) the two sets of weights were in close agreement, and (b) the α s based on the polychoric correlations were slightly larger than those based on the product-moment correlations. The most accurate item was the same in both cases (Item 15) with an

Table 1. Calibration Results. [AQ: 8]

(a) Goodness of fit and appropriateness of the unidimensional factor analysis model.		
Index	Continuous responses value (95% confidence interval)	Graded responses value (95% confidence interval)
RMSEA	0.0450 [0.0445, 0.0456]	0.0497 [0.0492, 0.0498]
Ts-RMSEA	0.072 (fair)	—
CFI	0.984 [0.983, 0.987]	0.987 [0.986, 0.990]
Ts-CFI	0.958 (close)	—
GFI	0.981 [0.980, 0.983]	0.980 [0.978, 0.983]
z-RMSR	0.054 [0.053, 0.055]	0.061 [0.060, 0.063]
ECV	0.90 [0.894, 0.913]	0.90 [0.894, 0.913]
G-H	0.95	0.96
(b) LRT results		
average Δ_i	0.20	0.43
Q (df)	3,262.3 (760)	2,945.9 (760)
Q-z	36.80	32.59

Note. RMSEA = root mean square error of approximation; Ts-RMSEA = T-size root mean square error of approximation; CFI = comparative fit index; Ts-CFI = T-size comparative fit index; GFI = goodness of fit index; z-RMSR = root mean square of residuals; ECV = explained common variance (ECV measures closeness to unidimensionality); G-H = generalized H index (G-H measures strength and replicability of the solution); LRT = likelihood ratio test.

Table 2. Reliability of the Person Estimates.

Estimate	DTCRM		DTGRM	
	ρ -PSD	ρ -S-H	ρ -PSD	ρ -S-H
EAP (θ)	0.95	0.95	0.94	0.94
EAP (σ^2)	0.80	0.81	0.82	0.80

Note. ρ -PSD = PSD-based marginal reliability; ρ -S-H = split-half reliability; EAP = expected a posteriori.

estimated α of .70 (DTCRM) and .74 (DTGRM). By using this item as a marker, $E(\sigma_i^2)$ was estimated at values of 1.06 (DTCRM) and 0.82 (DTGRM). Next, relative to this scaling, the IDD_s were estimated according to Equation 15.

As for the locations, the range of β_j values was $(-0.87, 1.40)$ in the DTCRM and $(-0.72, 1.32)$ in the DTGRM. In both cases, they were evenly distributed around 0, with means of 0.14 (DTCRM) and 0.17 (DTGRM). The correlation between both sets of estimates was 0.995.

EAP person estimates for the DTCRM and the DTGRM were obtained next. In both cases, the prior for θ was standard normal and the prior for σ^2 was inverse χ^2 with Scaling Parameter 3, and five degrees of freedom (see the appendix). Table 2 shows a summary of the accuracy of the estimates based on the marginal reliabilities in Equation 18 as well as on empirical split-half estimates.

To sum up, there is a high degree of agreement between (a) the results obtained from the DTCRM and the DTGRM, and (b) the PSD-based and the empirical split-half reliability estimates. In both models, the reliabilities of the person locations are those expected in a good

personality test, whereas those of the individual σ^2 estimates are lower but would be acceptable for many purposes. Finally, the product-moment correlations between the estimates produced by both models were 0.98 for the central locations and 0.86 for the PDDs.

The results of two respondents based on the DTGRM are now compared to illustrate the role of the PDD in individual assessment. The person location estimate of respondent no. 704 was $\hat{\theta}_{704} = 0.14$, whereas that of respondent no. 382 was $\hat{\theta}_{382} = 0.11$. So, in both cases, their estimated anxiety level was medium and about the same. However, the σ^2 estimate for respondent no. 704 was 5.51 while for respondent no. 382 it was 0.32, which shows that the second respondent answered the CTAC items much more consistently and precisely than the first. In accordance with this result, the PSD θ estimates were 0.42 for respondent no. 704 and 0.15 for respondent no. 382. The corresponding reliabilities in Equation 17 were 0.82 and 0.98. And, finally, the resulting confidence bands or 68% confidence intervals (i.e., $\hat{\theta} \pm \text{PSD}(\hat{\theta})$) were $[-0.28, 0.56]$ for respondent no. 704, and $[-0.04, 0.26]$ for respondent no. 382. Clearly, stronger inferences can be drawn on the basis of the person location estimate for the second respondent.

Finally, the role of PDD is illustrated as a moderator in prediction. Because no external variable was available, the “internal” split-half schema mentioned above was used: The raw scores in the first half were taken as the “predictor” and the raw scores in the second half as the “criterion.” Moderated Multiple Regression (e.g., Baron & Kenny, 1986) results showed that the moderating effects of the PDDs were significant: The F statistic value used to judge the increment of R^2 was 26.98 with df values of 1 and 757. The post hoc analysis was as follows: two subgroups were formed using Cureton’s (1957) 27% rule. The upper group contained the 205 respondents with the lowest PDDs (i.e., the most discriminating respondents), and the lower group contained the 205 with the highest (i.e., the least discriminating respondents). For the upper group, the split-half correlation was $r = .94$. For the lower group it was $r = .75$. Overall, the results suggest that the PDD estimates are useful in moderate prediction and are in the expected direction: The validity relations are stronger in the subgroups with the most discriminating respondents.

Discussion

Conventional psychometric modeling of typical-response measures considers lack of item discrimination as the sole source of measurement error. An alternative view stated by Lumsden (1978) was that items “are perfectly reliable” and that within-person variability (i.e., PDD) is the sole source of error. The view taken here is that both items and persons are sources of error and that the amount of error generally varies over persons and over items. This flexible scenario is thought to be the most plausible one for measurement in this domain. The problem, however, is that an analytically tractable modeling of this type does not exist at present for the most common item formats.

This article proposes a comprehensive approach to fitting dual models to continuous, graded, and binary item scores. The UVA on which it is based allows for a unified treatment of the models, so in all cases item calibration is performed via a FA of the inter-item correlation matrix with some additional restrictions. Simplicity and feasibility are possibly the main advantages of the proposal, as calibration is only slightly more complex than in standard models, whereas EAP scoring is quite similar to scoring a bidimensional model with independent factors.

The present proposal is a new, wide-scope development, and, as such, there are many points that require further research. At the methodological level, further simulation studies are needed mainly to establish the minimal conditions under which the dual models are expected to work well and provide reasonably accurate estimates for most of the individuals. Furthermore, future improvements in the approach proposed could be envisaged. For example, standard errors for

the parameter estimates obtained under constraints, mainly $E(\sigma_i^2)$, could be obtained by using the delta method or resampling procedures. Also, given the limitations of the LR statistic proposed, appropriate cutoff values should be determined, and the best option in future developments may be to obtain them via simulation. Finally, and more generally, theoretically superior procedures such as full-information estimation could be considered in the future for the proposed models. These procedures are far more complex than those proposed here and require the numerical integration of high dimensional integrals. Furthermore, it is not clear that in practice they will be clearly superior to the present proposal. However, the additional information they use from the data might avoid some of the constraints required in the limited-information case. In any case, the present proposal is fully open to improvements in the future.

Experience suggests that proposals such as the present can be used in practice only if they are implemented in widely available (and preferably free) programs. At present, work is on progress on an R program that will implement all the procedures proposed here, and which, hopefully, will soon be available for interested readers.

Appendix

Technical Details

Consider a test made up of $j = 1, 2, \dots, n$ items, and let \mathbf{x}_i be the full vector of responses given by individual i . The generic expression $P(X_j|\theta_i, \sigma_i^2)$ is used to denote the conditional probability (discrete case) or conditional density (continuous case) assigned to a specific item score for fixed θ_i and σ_i^2 . For the DTCRM, the conditional density is normal, with mean and variance given in Equation 3. For the DTBRM and the DTGRM, the conditional probabilities are those given in Equations 13 and 14, respectively.

The likelihood of \mathbf{x}_i can then be written generically as

$$L(\mathbf{x}_i|\theta, \sigma^2) = \prod_{j=1}^n P(X_{ij}|\theta, \sigma^2). \quad (21)$$

For a person parameter ξ ($\xi = \theta$ or σ^2), the expected a posteriori (EAP) point estimate is the mean of the posterior distribution of ξ given the respondent's item response pattern

$$\text{EAP} = \hat{\xi}_i = E(\xi|\mathbf{x}_i) = \frac{\int_{\theta} \int_{\sigma^2} \xi L(\mathbf{x}_i|\theta, \sigma^2) g(\theta) h(\sigma^2) d\sigma^2 d\theta}{\int_{\theta} \int_{\sigma^2} L(\mathbf{x}_i|\theta, \sigma^2) g(\theta) h(\sigma^2) d\sigma^2 d\theta}. \quad (22)$$

In all the models considered here, θ_i and σ_i^2 are assumed to be independent. So their joint distribution is the product $g(\theta)h(\sigma^2)$. As stated in the article, by default, $g(\theta)$ is taken as standard normal while σ_i^2 is scaled inverse χ^2 with d degrees of freedom and scaling parameter t . The mean and variance of this latter distribution are (Novick & Jackson, 1974)

$$E(\sigma_i^2) = \frac{t}{d-2}; \text{Var}(\sigma_i^2) = \frac{2t^2}{(d-2)^2(d-4)}. \quad (23)$$

The expectation in Equation 23 is obtained at the calibration stage according to Equation 15. To determine the parameters d and t , the simplest approach is to use a normal approximation and set a credibility interval (e.g., Swaminathan & Gifford, 1985). More specifically, experience with ML estimation in the DTCRM suggests that the expectations in Equation 23 are

usually close to 1, and that 0 to 5 is a reasonable range of values for the σ_i^2 estimates in most cases. Setting 1 as the mean and 5 as the upper end of a 95% credibility interval results in $d = 5$ and $t = 3$, which are the prior values chosen in the empirical study.

The double integral in Equation 22 can be approximated as accurately as required using numerical quadrature. More specifically, the applications described in the article used rectangular quadrature over $q = 40$ equally spaced points:

$$\text{EAP} \cong \frac{\sum_{k1=1}^q \sum_{k2=1}^q X_{km} L(\mathbf{x}_i | X_{k1}, X_{k2}, b, a) W(X_{k1}) W(X_{k2})}{\sum_{k1=1}^q \sum_{k2=1}^q L(\mathbf{x}_i | X_{k1}, X_{k2}, b, a) W(X_{k1}) W(X_{k2})}, \tag{24}$$

where $m = 1$ or 2 , X_{k1} and X_{k2} are the nodes and $W(X_{k1})$ and $W(X_{k2})$ are the weights for the one dimensional quadratures that approximate the distributions of θ and σ^2 , respectively.

The posterior standard deviation (PSD) is

$$\text{PSD}(\hat{\xi}_i) = \text{sqrt}\left(E(\xi^2 | \mathbf{x}_i) - \hat{\xi}_i^2\right). \tag{25}$$

The expectation of the squares in Equation 25 can be approximated by quadrature in the same form as in Equation 22.

As the number of items increases, the distribution of the θ EAP estimates approaches normality and the PSDs become equivalent to asymptotic standard errors (Bock & Mislevy, 1982). So, for a test of reasonable length, a normal-based confidence interval approach (strictly speaking, a credibility interval) for the EAP person location estimate of individual i can be constructed as

$$\hat{\theta}_i \pm z_c \text{PSD}(\hat{\theta}_i). \tag{26}$$

For both θ and σ^2 , explicit, approximate expression for the PSD can be obtained by considering that the information provided by the prior is numerically equivalent to an additional item to which all of the members of the population respond identically. For a test of reasonable length, it then follows that (see Wainer & Mislevy, 2000)

$$\frac{1}{\text{PSD}^2(\xi)} \cong I(\xi) - \frac{\partial^2 \log f(\xi)}{\partial \xi^2}, \tag{27}$$

where $I(\xi)$ is in this case the corresponding diagonal element of the expected information matrix, obtained as (e.g., Kendall & Stuart, 1977)

$$I(\xi) = -E\left[\frac{\partial^2 \log L(\theta, \sigma^2)}{\partial \xi^2}\right]. \tag{28}$$

The second term in Equation 27 is the amount of information contributed by the prior. For the case of θ , the contribution is simply 1. As for σ^2 , if the prior is inverse $\chi^2(d, t)$, then it follows that

$$-\left[\frac{\partial^2 \log f(\sigma^2)}{\partial (\sigma^2)^2}\right] = \frac{1}{(\sigma^2)^2} \left[\frac{t}{\sigma^2} - \frac{d+2}{2}\right]. \tag{29}$$

The approximate analytical expressions are then

$$\begin{aligned}\frac{1}{\text{PSD}^2(\hat{\theta}_i)} &\cong 1 + \sum_j^n \frac{1}{\sigma_i^2 + \sigma_{\epsilon j}^2} \\ \frac{1}{\text{PSD}^2(\sigma_i^2)} &\cong \left[\frac{1}{2} \sum_j^n \frac{1}{(\sigma_i^2 + \sigma_{\epsilon j}^2)^2} \right] + \frac{1}{(\sigma_i^2)^2} \left[\frac{t}{\sigma_i^2} - \frac{d+2}{2} \right].\end{aligned}\quad (30)$$

For the DTCRM,

$$\begin{aligned}\frac{1}{\text{PSD}^2(\hat{\theta}_i)} &\cong 1 + \sum_j^n \left(\frac{1}{\sigma_i^2 + \sigma_{\epsilon j}^2} \right) \frac{\phi^2 \left[\left(\frac{(\hat{\theta}_i - \delta_j)}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \right) \right]}{P_{ij} Q_{ij}} \\ \frac{1}{\text{PSD}^2(\sigma_i^2)} &\cong \left[\sum_j^n \phi^2 \left[\left(\frac{(\hat{\theta}_i - \delta_j)}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \right) \right] \frac{(\hat{\theta}_i - \delta_j)^2}{4(\sigma_i^2 + \sigma_{\epsilon j}^2)^3} \right] + \frac{1}{(\sigma_i^2)^2} \left[\frac{t}{\sigma_i^2} - \frac{d+2}{2} \right].\end{aligned}\quad (31)$$

For the DTBRM, where ϕ is the density of the standard normal distribution, $P_{ij} = P(X_j | \theta_i, \sigma_i^2)$, and $Q = 1 - P$. And, finally,

$$\begin{aligned}\frac{1}{\text{PSD}^2(\hat{\theta}_i)} &\cong 1 + \sum_j^n \left(\frac{1}{\sigma_i^2 + \sigma_{\epsilon j}^2} \right) \sum_k^c \frac{1}{P_{ijk}} \left[\left[\phi \left(\frac{(\hat{\theta}_i - \delta_{j,k-1})}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \right) - \phi \left(\frac{(\hat{\theta}_i - \delta_{j,k})}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \right) \right]^2 \right] \\ \frac{1}{\text{PSD}^2(\sigma_i^2)} &\cong \left[\sum_j^n \sum_k^c \frac{1}{P_{ijk}} \left[\phi \left[\frac{(\hat{\theta}_i - \delta_{j,k})}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \right] \frac{(\hat{\theta}_i - \delta_{j,k})}{2(\sigma_i^2 + \sigma_{\epsilon j}^2)^{3/2}} \right] - \left[\phi \left[\frac{(\hat{\theta}_i - \delta_{j,k})}{\sqrt{\sigma_i^2 + \sigma_{\epsilon j}^2}} \right] \frac{(\hat{\theta}_i - \delta_{j,k})}{2(\sigma_i^2 + \sigma_{\epsilon j}^2)^{3/2}} \right] \right]^2 \\ &\quad + \frac{1}{(\sigma_i^2)^2} \left[\frac{t}{\sigma_i^2} - \frac{d+2}{2} \right].\end{aligned}\quad (32)$$

For the DTGRM[AQ: 9]

Simulation Studies

Study 1: Item calibration and individual scoring with the DTGRM

Random samples of $N = 200$, $N = 500$, and $N = 1,000$ simulated responses were generated according to the DTGRM for (a) two test lengths, $n = 20$ and $n = 40$; and (b) two levels of IDs, $\sigma_\epsilon^2 = 1.55$ (which implies a common α value of 0.55) and $\sigma_\epsilon^2 = 0.29$ (which implies a common α value of 0.70) by using MATLAB programs written by the author. In all cases, 50 replicas per condition were used, the distribution of θ was standard normal, and the distribution of σ^2 was inverse chi-square with $t = 3$ and $d = 5$. The β item locations were uniformly distributed between -1.5 and 1.5 , and items were discretized into five response categories using Muthén and Kaplan's (1985) thresholds for obtaining centered distributions. It should be noted, however, that the combination of the chosen β values with the standard thresholds gave rise in some cases to quite skewed item distributions[AQ: 10]

First, the simulated responses were calibrated by fitting the unidimensional factor analysis (FA) model to the polychoric inter-item correlation matrices using unweighted least squares (ULS) estimation. The recovery of the generating parameters was assessed with the mean, bias, and root mean squared error (RMSE) of the standardized loadings. Goodness of fit was assessed with two statistics: the root mean squared residual (RMSR) and the goodness of fit index (GFI) (see McDonald, 1999). The results are in the upper panel of Table A1.

Results in Table A1(a) can be summarized as follows: In all cases, the goodness of fit statistics agree with the expectations derived from the null hypothesis of model-data fit, and the item parameters are reasonably well recovered. With small to medium sample sizes, the loading estimates are slightly attenuated, and the bias approaches zero as the sample size increases, which is reasonable.

The second part of the study is the scoring stage. For each simulee, EAP estimates of both θ and σ^2 were obtained by using rectangular quadrature over 40 equally spaced points and the correct prior distributions above. Thus, in this respect, the EAP-based results must be considered to have been obtained under “ideal” conditions. In addition to the bias and RMSE of both person parameters, the measures of accuracy in this case were (a) the product-moment correlation between the individual estimates and the corresponding true values, and (b) the marginal reliability estimates in Equation 18. Measure (a) can be interpreted as an index of reliability. So, if both types of measure agree, the marginal reliabilities must be approximately equal to the square of the reliability indices. The results are shown in the lower panel of Table A1.

Results in Table A1(b) generally behave according to the theoretical expectations. For both θ and σ^2 , the accuracy increases with test length and item discriminating power. Also for both parameters, the relations between both measures of accuracy agree reasonably well and the increases in the marginal reliabilities with test length tend to agree with the predictions obtained by using the Spearman–Brown formula.

In the conditions used in the study, the accuracy of the θ estimates is quite good in all cases. The accuracy of the σ^2 estimates, however, is clearly lower, as expected. Even so, the present results suggest that individual estimates of σ^2 , which are accurate enough for practical purposes, might be obtained in tests of only 20 items provided that the IDD is reasonably low.

Study 2: Behavior of the likelihood ratio test (LRT) Q statistic with the DTGRM and EAP estimates

The simulated data used in H_1 were the same as in Study 1 above, but only the condition $\alpha = .70$ was used. Under H_0 , random samples were generated with the same characteristics as those in H_1 except that all the simulees had the same constant PDD value which was the mean of σ^2 in H_1 . To avoid dependency of the results on sample size, the ratio Q/N was reported instead of Q . Results are in Table A2.

Results in Table A2 show that the statistic allows the correct model to be distinguished in all conditions and that power increases with test length, as expected. However, the statistic is conservative, and, under H_0 , it provides values systematically smaller than the chi-square expectations, especially in small samples.

New References Used in the Appendix

- Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 2). London, England: Charles Griffin.
- McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

Table A1. Results of the Simulation Study [AQ: 11]

Model size	20						40					
	200		500		1,000		200		500		1,000	
	$\alpha = .55$	$\alpha = .70$	$\alpha = .55$	$\alpha = .70$	$\alpha = .55$	$\alpha = .70$	$\alpha = .55$	$\alpha = .70$	$\alpha = .55$	$\alpha = .70$	$\alpha = .55$	$\alpha = .70$
(a) Item calibration												
$M\alpha$	0.52	0.68	0.54	0.68	0.55	0.70	0.52	0.68	0.54	0.68	0.55	0.70
Bias α	-0.03	-0.02	-0.01	-0.02	0.00	0.00	-0.03	-0.02	-0.01	-0.02	0.00	0.00
RMSE α	0.03	0.03	0.01	0.02	0.01	0.01	0.03	0.02	0.01	0.02	0.01	0.01
GFI	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	(0.01)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
z-RMSR	0.04	0.04	0.03	0.03	0.02	0.02	0.04	0.05	0.04	0.03	0.02	0.02
	(0.003)	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
(b) Scoring												
$r(\hat{\theta}, \theta)$	0.94	0.97	0.94	0.97	0.94	0.97	0.97	0.98	0.97	0.98	0.97	0.98
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$r(\sigma^2, \hat{\sigma}^2)$	0.63	0.80	0.62	0.79	0.62	0.79	0.75	0.88	0.75	0.87	0.76	0.87
	(0.12)	(0.08)	(0.07)	(0.05)	(0.07)	(0.04)	(0.10)	(0.04)	(0.08)	(0.03)	(0.06)	(0.03)
$\rho(\hat{\theta}, \hat{\theta})$	0.88	0.93	0.88	0.94	0.88	0.94	0.93	0.96	0.93	0.97	0.93	0.97
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$\rho(\sigma^2, \hat{\sigma}^2)$	0.51	0.64	0.50	0.68	0.51	0.68	0.63	0.79	0.64	0.80	0.64	0.80
	(0.04)	(0.02)	(0.03)	(0.04)	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.02)	(0.01)	(0.01)
Bias $\hat{\theta}$	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RMSE $\hat{\theta}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Bias $\hat{\sigma}^2$	-0.01	-0.01	-0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RMSE $\hat{\sigma}^2$	0.03	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Note. RMSE = root mean square error; GFI = Goodness of Fit Index; z-RMSR = Root Mean Square of Residuals.

Table A2. Simulation Results for the LRT Statistic.

Model size	20-item			40-item		
	Sample size	200	500	1,000	200	500
Constant GRM (H_0) Q/N	0.56 (0.03)	0.65 (0.05)	0.65 (0.04)	0.37 (0.06)	0.80 (0.05)	0.80 (0.04)
DTGRM (H_1) Q/N	1.26 (0.26)	1.51 (0.30)	1.50 (0.10)	2.63 (0.20)	2.55 (0.20)	2.56 (0.15)

Note. LRT = likelihood ratio test; GRM = graded-response model.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349-364.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 61-101). Mahwah, NJ: Lawrence Erlbaum.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has been possible with the support of Ministerio de Economía, Industria y Competitividad, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF) (PSI2017-82307-P).

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.
- Brown, A., & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307-333). New York, NY: Routledge.
- Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*, 201-225.
- Cureton, E. E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika*, *22*, 293-296.
- DeFleur, M. L., & Catton, W. R. (1957). The limits of determinacy in attitude measurement. *Social Forces*, *35*, 295-300.
- Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research*, *37*, 521-542.
- Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement*, *28*, 126-140.
- Ferrando, P. J. (2007). A Pearson-type-VII item response model for assessing person fluctuation. *Psychometrika*, *72*, 25-41.
- Ferrando, P. J. (2009). A graded response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology*, *62*, 641-662.

- Ferrando, P. J. (2013). A general linear framework for modeling continuous responses with error in persons and items. *Methodology*, *9*, 150-161.
- Ferrando, P. J. (2014). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behavioral Research*, *49*, 390-405.
- Ferrando, P. J. (2016). An IRT modeling approach for assessing item and person discrimination in binary personality responses. *Applied Psychological Measurement*, *40*, 218-232.
- Fiske, D. W. (1968). Items and persons: Formal duals and psychological differences. *Multivariate Behavioral Research*, *3*, 393-401.
- LaHuis, D. M., Barnes, T., Hakoyama, S., Blackmore, C., & Hartman, M. J. (2017). Measuring traitedness with person reliabilities parameters. *Personality and Individual Differences*, *109*, 111-116.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, *4*, 269-290.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Applied Psychological Measurement*, *37*, 497-498.
- Lubbe, D., & Schuster, C. (2016). Consistent differential discrimination model estimation. *Multivariate Behavioral Research*, *51*, 581-587.
- Lubbe, D., & Schuster, C. (2017). The graded response differential discrimination model accounting for extreme response style. *Multivariate Behavioral Research*, *52*, 616-629.
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, *31*, 19-26.
- Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement*, *4*, 1-7.
- Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, *35*, 63-78.
- McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, *6*, 379-396.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381. [AQ: 12]
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-195.
- Mosier, C. I. (1942). Psychophysics and mental test theory II: The constant process. *Psychological Review*, *48*, 235-249.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered, categorical and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.
- Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York, NY: McGraw-Hill.
- Pallero, R., Ferrando, P. J., & Lorenzo-Seva, U. (1998). Questionnaire Tarragona of anxiety for blind people. In E. Sifferman, M. Williams, & B. B. Blasch (Eds.), *The 9th international mobility conference proceedings* (pp. 250-253). Atlanta, GA: Rehabilitation Research and Development Center. [AQ: 13]
- Paunonen, S. V. (1988). Trait relevance and the differential predictability of behavior. *Journal of Personality*, *56*, 599-619.
- Reise, S. P., & Waller, N. G. (1993). Traitdness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143-151.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Iowa City, Iowa: Psychometric Society.
- Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement*, *11*, 355-370.
- Taylor, J. B. (1977). Item homogeneity, scale reliability, and the self-concept hypothesis. *Educational and Psychological Measurement*, *37*, 349-361.
- Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 622-663.
- Torgerson, W. (1958). *Theory and methods of scaling*. New York, NY: Wiley.

-
- van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339-356.
- Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 319-330.