Article

Applied Psychological Measurement
2016, Vol. 40(3) 218–232

DOI: 10.1177/0146621615622633
apm.sagepub.com
SAGE

An IRT Modeling Approach for Assessing Item and Person Discrimination in Binary Personality Responses

Pere J. Ferrando[1]

Abstract

Conventional item response theory (IRT) modeling of personality responses considers two item characteristics—location and discrimination—but only one person characteristic—location or trait level. An IRT modeling approach that jointly considers item and person discriminations, however, is thought to be more realistic and appropriate in this domain and has several potential advantages. This article develops a model of this type for unidimensional binary responses together with procedures for estimating item and person parameters and assessing model appropriateness (including person fit). A series of preliminary simulations suggests that the approach is feasible, and a real-data example illustrates the potential advantages with respect to the standard two-parameter model. Limitations of the proposal and further work are also discussed.

Keywords

personality measurement, item response theory, person discrimination, person reliability, person fit, EAP estimation, likelihood statistics

Fiske (1968) discussed the psychometric analysis of personality data and considered that each characteristic or parameter assessed at the item level has a counterpart or ''dual'' at the person level. In the case of unidimensional binary responses dealt with here, the person dual of the item location is the individual trait level (i.e., the person location). The item discrimination, however, reflects the accuracy of the item as a measure of the trait (Lord & Novick, 1968). Therefore, the corresponding dual is a person parameter that reflects the accuracy of the individual as a respondent. A parameter of this type is consistent with evidence: Some individuals respond in a highly accurate, almost deterministic way, whereas others respond much more randomly (Ferrando, 2004, 2013; Fiske, 1968; Guilford, 1959; Voyce & Jackson, 1977).

Fiske (1968) and Voyce and Jackson (1977), among others, proposed classical-test-theory-based analyses in which both item and person discrimination were assessed (for an historical review, see Ferrando, 2004). Also, in the context of person-fit analysis, ''dual'' proposals have been made based on an item response theory (IRT) framework (Levine & Rubin, 1979; Reise, 2000; Reise & Due, 1991). At a more specific level, however, the IRT models commonly used

[1]Rovira i Virgili University, Tarragona, Spain

Corresponding Author:
Pere J. Ferrando, Facultad de Psicologia, Rovira i Virgili University, Carretera Valls s/n, 43007 Tarragona, Spain.
Email: perejoan.ferrando@urv.cat

in personality measurement do not include both item and person discrimination parameters. The only models that do appear to be the two-parameter 3 model (2P3M) described by Lumsden (1980) and the generalized multiplicative model (GMM) proposed by Strandmark and Linn (1987).

The starting point of this article is that a dual model that jointly considers location and discrimination parameters for both items and individuals is ''a priori'' the more realistic way of modeling personality responses within an IRT framework (Ferrando, 2013; Lumsden, 1980). At the same time, however, a model of this type would be of little value in practice if it were too complex and/or unable to provide accurate and stable parameter estimates. These limitations are certainly applicable to the two proposals mentioned above. Estimation of the 2P3M at the individual level is considered unfeasible except in the linear-continuous case (Ferrando, 2013), and, for binary responses, only a restricted version with constant item discriminations has been proposed (Ferrando, 2004, 2007). As for the GMM, Strandmark and Linn (1987) proposed a complex joint iterative procedure that did not guarantee the correctness and stability of the estimates. Thus, it is only to be expected that the GMM has not been used to date in any application but only as a source model for generating person misfit (Reise & Due, 1991).

The aim of this article is to propose a simple and mathematically tractable IRT dual model intended for personality binary items together with procedures for (a) fitting it and obtaining item and person estimates, (b) assessing model appropriateness (including person fit), and (c) interpreting the person estimates and using them in applications. The basic model is a modified version of the GMM that includes a series of results previously reported by Ferrando (2004, 2007). So, although some new results have been obtained, the ultimate purpose of the proposal is to provide a workable approach that can be used in applied personality assessment.

Description of the Model and Main Results

Consider a personality test, made up of n binary items, that measures a trait θ. Let X_{ij} (0 or 1) be the observed score of respondent i on item j. In the standard two-parameter model (2PM), the expected score of respondent i on item j, which is also the probability that respondent i would endorse item j, is given by

$$P_j(\theta_i) = P\left(X_{ij} = 1 | \theta_i, b_j, a_j\right) = \Phi\left(a_j\left(\theta_i - b_j\right)\right) \cong \Psi\left(Da_j\left(\theta_i - b_j\right)\right), \quad (1)$$

where Φ is the Cumulative distribution function (c.d.f.) of the standard normal distribution, Ψ is the logistic function, and $D = 1.702$. The parameters θ_i and b_j are, respectively, the person and item locations on the trait continuum, and for identification purposes, θ is assumed to be distributed with zero mean and unit variance in the population of interest. The parameter a_j is considered to be positive and is the item discrimination. The difference $\theta_i - b_j$ is the signed person–item distance (PID; Kuncel, 1973; Voyce & Jackson, 1977) and is the primary response determinant. So, when the person location dominates the item location (i.e., $\theta_i > b_j$) the expected score is above 0.5 (i.e., the Response Scale midpoint). This dominance mechanism is thought to be an appropriate way to model the responding process that takes place when many binary personality items are answered (Kuncel, 1973; Voyce & Jackson, 1977). However, it is not claimed to be universally valid for all of them. Thus, neutral items that contain ''average'' modifiers or use double-barreled statements might operate more consistently with a proximity mechanism (e.g., Cao, Drasgow, & Cho, 2015).

The model proposed in this article, named *D2PMM* (dual two-parameter multiplicative model), is an extension of Equation 1, and its basic equation is

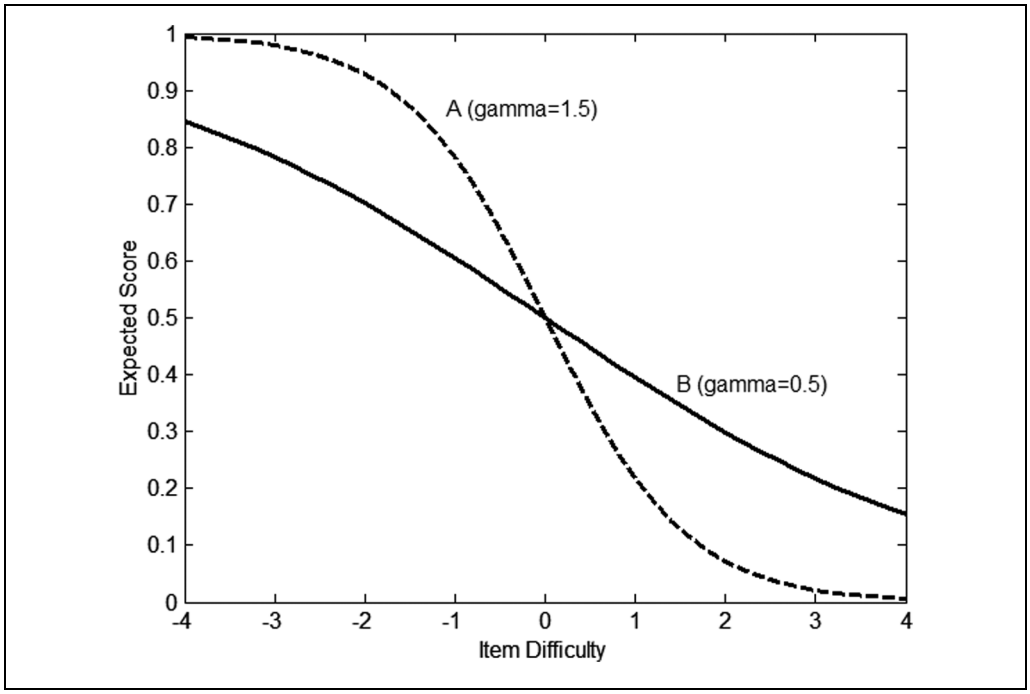

Figure 1. Expected score as a function of item difficulty for two respondents with different person discrimination.

$$P_j(\theta_i, \gamma_i) = P\left(X_{ij} = 1 | \theta_i, \gamma_i, b_j, a_j\right) = \Phi\left(\gamma_i a_j \left(\theta_i - b_j\right)\right). \quad (2)$$

Equation 2 is essentially the same as that proposed by Strandmark and Linn (1987) except that they proposed a logistic version. The normal-ogive choice proved to be more appropriate and treatable than the logistic version for the new developments described below.

The person parameter γ_i in Equation 2 is also assumed to be positive, reflects the discriminating power of the individual, and is the dual of a_j. Furthermore, the item and person discriminations are assumed to be independent (Lumsden, 1980). The product $\gamma_i a_j$ moderates the sensitivity of the responding process to the PID primary determinant. So, when both γ_i and a_j are high, the responding process becomes more deterministic and approaches a Guttman process. When either γ_i or a_j or both approach 0, the process becomes more and more random, and the expected score approaches 0.5 no matter what the PID is.

If the expected item scores derived from Equation 2 were plotted as a function of item location for a fixed respondent, the D2PMM person response function (PRF) of this respondent would be obtained. The D2PMM-PRF, which is shown in Figure 1, is a decreasing two-parameter ogive whose slope is proportional to the person discrimination γ_i in which the person location θ_i defines the point along the b continuum at which the expected score is 0.5.

The item response function (IRF) of the D2PMM is next defined as the expectation of Equation 2 over γ. A conceptual interpretation is that the IRF is the expected score on item j for the sub-population of respondents with the same location θ_i. An alternative, but equivalent, interpretation is that the IRF is the probability of endorsing item j for a random respondent with location θ_i (see Ferrando, 2004, 2007).

$$IRF = P_j(\theta_i) = E_\gamma\left(P_j(\theta_i,\gamma)\right) = E_\gamma\left[\Phi\left(\gamma a_j\left(\theta_i - b_j\right)\right)\right]$$
$$= \int_\gamma \Phi\left(\gamma a_j\left(\theta_i - b_j\right)\right) f(\gamma) d\gamma. \quad (3)$$

Because γ can be viewed as a slope parameter that is bounded away from 0, a reasonable and mathematically convenient choice for $f(\gamma)$ in Equation 3 is the scaled-chi (χ) distribution (also known as the generalized Rayleigh distribution), which has the parameters ν (degrees of freedom) and λ (scale parameter; see Swaminathan & Gifford, 1985), and whose probability density function (p.d.f.) is proportional to

$$f(\gamma) \propto \gamma^{\nu-1} \exp\left(-\frac{1}{2}\left(\lambda\gamma^2\right)\right). \quad (4)$$

Using Equation 4 in Equation 3, the IRF is found to be (see Ferrando, 2007)

$$IRF = P_j(\theta_i) = F_{t,\nu}\left(\sqrt{\frac{\nu}{\lambda}} a_j\left(\theta_i - b_j\right)\right) \cong \Phi\left(E(\gamma) a_j\left(\theta_i - b_j\right)\right), \quad (5)$$

where $F_{t,\nu}$ is the c.d.f. of Student's distribution with ν degrees of freedom. The IRF, then, is a Student's ogive with slope parameter $\sqrt{(\nu/\lambda)}a_j$. As ν increases, the curve approaches a normal ogive, and $\sqrt{\nu/\lambda}$ approaches $E(\gamma)$ (see Ferrando, 2007). Even at small ν, however, both IRFs are close and, in fact, Ferrando (2007) found that estimation of the Student ogive was generally unstable because of the difficulty involved in distinguishing it from the normal ogive. For this reason, the limiting normal ogive on the right-hand side of Equation 5 will be taken as the IRF of the D2PMM. The slope of this IRF is the product $E(\gamma)a_j$ and cannot be identified without further restrictions. The constraint proposed here is to set $\nu = \lambda$, which, for interpretive purposes, makes $E(\gamma)$ approximately 1 (see Equation 14 in the online appendix).

The result discussed above suggests that in the calibration stage, the item location and item slope parameters of the D2PMM could be estimated by fitting the standard normal-ogive 2PM to the matrix of item scores. However, the fact that the 2PM IRF closely approximates the D2PMM at the single-item level does not in itself guarantee that the approximation, which is proposed, will be generally good enough.

As discussed below, in this article, the author proposes a conventional two-stage estimation procedure (i.e., calibration and scoring) in which items are first calibrated by using the limited-information unweighted-least-squares procedure proposed by McDonald (e.g., McDonald, 1997). This calibration procedure is chosen because it is quite robust and makes satisfactory estimates even when tests are long and sample sizes not too large (the situation expected in most applications). Now, the input in McDonald's procedure is the inter-item cross-product matrix. So, in the calibration stage, the 2P normal-ogive item estimates can be expected to be reasonably close to the D2PMM item parameters if (a) the conditional and marginal proportions of single-item endorsements implied by both models are close (see above) and (b) the conditional and marginal joint probability of endorsement for pairs of items are also close enough.

The closeness of approximation regarding point (b) above was assessed analytically and computationally and is discussed in detail in the online appendix (Section 2, Equations 20-25). The results show that (a) for the conditions that are expected in normal-range personality measurement and (b) for a plausible, non-extreme chi-distribution for γ (say $\nu = \lambda = 5$ or more), the normal-ogive approximation is good even for extreme items.

To sum up, the D2PMM proposed in this section is essentially a 2P3M-type model according to Lumsden's (1980) taxonomy. At the individual level, its PRF is a two-parameter normal

ogive. At the item level, its IRF is expected to be well approximated by the IRF of the 2P normal-ogive model. The main differences between both models lie in their PRFs: In the standard 2PM, only the person location θ_i is considered, so the implicit PRF is a one-parameter normal ogive with the same slope for all the respondents. In the limiting case in which all the respondents have the same amount of discrimination, the D2PMM will reduce to the standard 2P normal-ogive model.

Relation With Other Approaches

Moustaki and Knott (2000) proposed a family of generalized latent trait models characterized by (a) a link function, (b) a random component, and (c) a systematic component. The D2PMM can be considered as a model belonging to this family in which the random component is the Bernoulli distribution of the 0 to 1 scores, the link function is the probit function (Φ^{-1}), and the systematic component is the term $\gamma_i a_j(\theta_i - b_j)$ (see Equation 2). Furthermore, the D2PMM can be related to two existing classes of factor analytic (FA) models by re-expressing the systematic component term in the two ways discussed below.

Consider first the following re-expression derived from the basic Equation 2:

$$\Phi^{-1}\left(P_j(\theta_i,\gamma_i)\right) = \gamma_i a_j\left(\theta_i - b_j\right) = \left(-\gamma_i a_j b_j\right) + \left(\gamma_i a_j\right)\theta_i = \mu_{ij} + \lambda_{ij}\theta_i. \quad (6)$$

With this formulation, the systematic component of the D2PMM is that of a linear heterogeneous FA model in which both the intercepts and the weights vary over individuals. FA models of this type have been proposed mainly for the linear case in which the link function is the identity function (e.g., Ansari, Jedidi, & Dube, 2002; Ferrando, 2013; Kelderman & Molenaar, 2007).

The second re-expression is

$$\Phi^{-1}\left(P_j(\theta_i,\gamma_i)\right) = \left(-a_j b_j\right)\gamma_i + \left(a_j\right)\gamma_i\theta_i = \lambda_{j1}\gamma_i + \lambda_{j2}\gamma_i\theta_i. \quad (7)$$

Expressed as in Equation 7, the systematic component becomes that of a regression model with two orthogonal latent variables (γ and θ), a linear main-effect term, and an interaction term expressed as the product of the latent variables. For the continuous case with identity link function, models of this type have been considered in the structural equation modeling literature (e.g., Cudeck, Harring, & du Toit, 2009). Furthermore, Rizopoulos and Moustaki (2008) have considered a logit-link model with two orthogonal factors, two main-effect terms, and a product term, which is more closely related to Equation 7. The D2PMM, however, differs from the standard formulation of interactive models in two ways. First, Model 7 has only one main-effect term. Second, and more important, the standard formulation of interactive models (including Rizopoulos and Moustaki's) is an FA formulation in which both latent variables are modeled as common factors with the same distribution and metric (generally standard normal). In contrast, in the D2PMM, the common factor is a standard variable centered at 0, whereas the individual discrimination variable is a positive variable distributed asymmetrically. The interaction effect in the D2PMM, then, is not ''symmetric.'' Rather, the role of γ is to amplify or reduce the effect of θ on the item responses.

Parameter Estimation and Properties of the Person Estimates

Rizopoulos and Moustaki (2008) proposed a full-information maximum-likelihood estimation procedure for generalized latent variable models that could be adapted to fit the D2PMM. The

procedure enjoys good statistical properties but is complex. So, given that the present proposal aims for simplicity and practicality, the author proposes, instead, to use a far simpler method that follows a conventional two-stage approach (i.e., calibration and scoring). In the calibration stage, the 2P normal-ogive model is fitted to the inter-item cross-product matrix by using the limited-information unweighted-least-squares procedure developed by McDonald (1997) and the free software that is available for this approach (Fraser & McDonald, 2012). If the D2PMM is correct, given the results discussed above, the location and discrimination item estimates obtained with this approach are expected to be good estimates of their corresponding D2PMM parameters (this point is further assessed in the simulation study below). Next, provided that the fit is acceptable, the item parameter estimates are taken as fixed and known and used in the scoring stage, which is based on Equation 2. A calibrated set of items can also be used if available.

The scoring procedure—Bayes expected a posteriori (EAP, Bock & Mislevy, 1982)—is selected for two main reasons. First, to ensure that the person estimates (especially γ) fall within reasonable values and second, to use the information available at the calibration stage (mainly that the distribution of γ is χ with approximately unit expectation). The prior for θ will usually be taken as standard normal, but other specifications are possible. As for γ, the prior can be fully specified by setting a credibility interval and then solving for the variance parameter (see Ferrando, 2007; Swaminathan & Gifford, 1985).

EAP estimation of θ and γ in the D2PMM computed by quadrature is conventional and is detailed in the online appendix. The output consists of the EAP point estimates and the posterior standard deviations (*PSD*s), which serve as standard errors (e.g., Bock & Mislevy, 1982). A *PSD*-based reliability estimate can further be obtained as

$$\rho(\hat{\theta}_i) = 1 - \frac{PSD(\hat{\theta}_i)^2}{Var(\theta)}, \qquad \rho(\hat{\gamma}_i) = 1 - \frac{PSD(\hat{\gamma}_i)^2}{Var(\gamma)}. \tag{8}$$

The θ estimates obtained from the D2PMM have two properties that distinguish them from the standard 2PM. First, provided that the D2PMM is correct, they are unbiased. Second, they have a different degree of precision depending on the amount of person discrimination.

The first property can be illustrated graphically. Figure 1 shows the D2PMM PRFs for two hypothetical respondents—A and B—with the same trait level ($\theta = 0$) but different discriminations ($\gamma_A = 1.5$, and $\gamma_B = 0.5$). As noted by Lumsden (1977), if their trait levels are estimated by the standard 2PM, then the relative estimates will be biased by the location of the items. In a ''difficult'' test, the trait level of B will be seen as higher than A, and the opposite will occur in an ''easy'' test. In contrast, the D2PMM-based estimates, which take into account the differential discrimination of both respondents, are expected to be unbiased.

The second property has already been discussed by Reise and Due (1991) who noted that responses given by a less-discriminating individual provide less psychometric information for estimating θ. This point can be made more explicit here by using the following relation (detailed in the online appendix):

$$\frac{1}{PSD^2(\theta_i)} \cong 1 + \gamma_i^2 \sum_{j=1}^{n} a_j^2 \frac{\phi^2\left(\gamma_i a_j(\theta_i - b_j)\right)}{P_j(\theta_i, \gamma_i)\left(1 - P_j(\theta_i, \gamma_i)\right)} = 1 + I(\theta_i), \tag{9}$$

where $I(\theta)$ is the element of the information matrix corresponding to θ (see the online appendix). Other factors being constant, $PSD(\theta)$ decreases (and the confidence interval becomes

Table 1. Expected Information per Unit Test Length for θ (Left) and γ (Right).

	$I(\theta)$					$I(\gamma)$				
	θ					θ				
γ	−2	−1	0	1	2	−2	−1	0	1	2
Item discrimination = 0.4										
0.50	0.023	0.024	0.025	0.024	0.023	0.587	0.364	0.284	0.364	0.587
0.75	0.047	0.051	0.053	0.051	0.047	0.442	0.312	0.261	0.312	0.442
1.00	0.072	0.084	0.088	0.084	0.072	0.291	0.247	0.228	0.247	0.291
1.50	0.123	0.152	0.162	0.152	0.123	0.110	0.132	0.149	0.132	0.110
2.00	0.179	0.219	0.228	0.219	0.179	0.049	0.070	0.083	0.070	0.049
Item discrimination = 0.9										
0.50	0.085	0.100	0.106	0.100	0.085	1.156	1.084	1.056	1.084	1.156
0.75	0.144	0.177	0.188	0.177	0.144	0.399	0.522	0.613	0.522	0.399
1.00	0.209	0.252	0.259	0.252	0.209	0.181	0.270	0.309	0.270	0.181
1.50	0.356	0.389	0.389	0.389	0.356	0.066	0.093	0.094	0.093	0.066
2.00	0.506	0.519	0.518	0.519	0.506	0.034	0.040	0.040	0.040	0.034

narrower) the more discriminating the individual is. Conceptually then, more confidence can be placed in the trait estimate of an accurate respondent than in that of a low-discriminating respondent.

In practice, the distinctive features just discussed can only be considered as advantages with respect to the standard 2PM if the person discrimination estimates have a minimal degree of precision. Otherwise, the bias correction and differential accuracy predictions might be misleading, and it might even be preferable to use the standard 2PM trait estimates.

The expression for $PSD(\theta)$ is (see the online appendix) as follows:

$$\frac{1}{PSD^2(\gamma_i)} \cong \left(\frac{\nu-1}{\gamma_i^2}+\lambda\right)+\sum_{j=1}^{n} a_j^2\left(\theta_i-b_j\right)^2 \frac{\phi^2\left(\gamma_i a_j\left(\theta_i-b_j\right)\right)}{P_j(\theta_i,\gamma_i)\left(1-P_j(\theta_i,\gamma_i)\right)} = \left(\frac{\nu-1}{\gamma_i^2}+\lambda\right)+I(\gamma_i). \tag{10}$$

According to Equation 10, the amount of measurement error in this case depends mainly on three factors (apart from the prior contribution): (a) the number of items, (b) the squared PIDs, and (c) the amount of item discrimination. So, for a single respondent with a given location θ_i, γ can only be estimated accurately if there are enough items with good discrimination that are sufficiently distant from this location. Overall, then, reliable estimation of γ requires a test that is long enough and made up of items that have a wide dispersion of locations across the trait continuum and a certain amount of discrimination.

For both θ and γ, the expected *PSD*s corresponding to different conditions can be computed analytically with no need for any empirical data (Ferrando, 2004). Table 1 displays the expected information per unit test length for θ (left) and γ (right). The values in the upper panel are computed for a medium-low average item discrimination of $a = 0.40$. Those in the lower panel are obtained for a medium-high value of $a = 0.90$. In both cases, the item locations are assumed to be uniformly distributed in the interval -3 to $+3$.

For a test of length n, the expected amount of information is obtained by multiplying the corresponding value in the table by n. Then, the corresponding *PSD*s and reliabilities are obtained

by using Equations 8, 9, and 10. For example, assume that a 40-item test with an average discrimination of 0.90 is administered to an ''average'' respondent with person estimates: $\theta = 0$ and $\gamma = 1$. So, according to Table 1, $I(\theta) = 40 \times 0.259 = 10.36$, and $I(\gamma) = 40 \times 0.309 = 12.36$. Assume next that the prior for θ is standard normal and the prior for γ is $\chi(3, 3)$, which leads to $\sigma(\gamma) = 0.41$. Then, according to Equations 9 and 10, the corresponding *PSD*s are $PSD(\theta) = 0.297$, and $PSD(\gamma) = 0.248$. Finally, according to Equation 8, the reliability of the estimates are $\rho(\theta) = 0.91$ and $\rho(\gamma) = 0.64$ (recall that $\sigma(\gamma) = 0.41$).

So far, person discrimination has been considered only as a means for improving θ estimation. However, it might be of interest beyond this auxiliary role. First, it might act as a moderator variable regarding the relation between trait estimates and relevant criteria (e.g., Culpepper, 2010). More specifically, if the precision of trait estimates depends on γ, then stronger validity relations are expected for highly discriminating individuals (e.g., Lord & Novick, 1968, Section 3.9). Second, γ might be an indicator of an individual-differences dimension of sensitivity to the normative ordering of the items that is known as ''person reliability'' (Ferrando, 2004; Guilford, 1959; Lumsden, 1977). This dimension would partly characterize the responding behavior of the individual (from almost random to almost deterministic) and might also have a theoretical foundation: It might be related to the strength, clarity, and degree of organization by which the trait is internally represented (e.g., Taylor, 1977; Tellegen, 1988). Furthermore, a dimension of this type might well be related to such personality variables as impulsivity, conscientiousness, conformity, or restraint (Donlon & Fischer, 1968; Ferrando, 2004, 2007).

Assessing Model Appropriateness

This section will discuss procedures for assessing model-data fit at the overall level and at the individual level (i.e., person fit). At the overall level, if the goodness-of-fit results obtained at the calibration stage are satisfactory, then both the D2PMM and the 2PM can be considered to be appropriate because they both predict essentially the same IRF. A key issue then is to assess whether the more flexible but more parameterized D2PMM is more appropriate than the simpler 2PM.

The incremental appropriateness of the D2PMM with respect to the 2PM can be assessed by using a likelihood ratio (LR) approach (e.g., Ferrando, 2013). Consider the response pattern of respondent i, and let (a) $L_i^1(\hat{\theta}_i, \hat{\gamma})$ be the value of the D2PMM likelihood function (see the online appendix) evaluated by using both person estimates and (b) $L_i^0(\hat{\theta}_i^{(f)}, 1)$ be the corresponding value obtained under the restriction that all the person discriminations have a fixed value of $\gamma = 1$ (i.e., the fixed γ mean). As mentioned above, with this restriction, the trait estimates $\hat{\theta}_i^{(f)}$ are those obtained with the standard 2PM. The LR statistic and its standard transformation are

$$\Lambda_i = \frac{L_i^0\left(\hat{\theta}_i^{(f)}, 1\right)}{L_i^1\left(\hat{\theta}_i, \hat{\gamma}_i\right)}, \qquad s_i = -2\ln(\Lambda_i) \tag{11}$$

The statistic Λi is a descriptive normed index with values in the range 0 to 1. Values close to 1 indicate that the simpler, more restricted 2PM is appropriate for this respondent. As for s_i, under very restrictive conditions, it could be considered as a value drawn at random from a χ^2 distribution with one degree of freedom. So, by assuming experimental independence between respondents, the sum $Q = \Sigma s_i$ would asymptotically approach a χ^2 distribution with N degrees of freedom (see Ferrando, 2013). This sum Q is proposed here as the overall index for assessing whether the D2PMM fits the data better than the standard 2PM. However, as discussed in

Ferrando (2013), Q should not be used as a strict inferential measure but rather in a more exploratory way in which the theoretical distribution is only a useful reference.

Regarding person-fit assessment, the parameter γ defines a continuum ranging from random responding to deterministic responding, and so a key assumption of the D2PMM is that $\gamma > 0$. An individual who gives responses that are opposite to the normative ordering of the item locations will produce a negative value of γ, so his or her pattern would be inconsistent with the D2PMM. To assess this type of inconsistency, a modified, conditional version of the personal biserial correlation (Donlon & Fischer, 1968) termed $c\text{-}r_{\text{perbis}}$ is proposed here as a person-fit statistic. The index is the biserial correlation between the response vector and the vector of expected item scores given the EAP trait estimate of the individual obtained from the standard 2PM. If the D2PMM holds, then $c\text{-}r_{\text{perbis}}$ is positively related to γ, its value must be near 0 for an almost-random pattern, and must approach 1 as the amount of discrimination of the individual increases. A negative value indicates that the individual's responses are not consistent with the model, so his or her estimates cannot be validly interpreted.

As a summary of the proposal discussed so far, the online appendix includes a guideline on how the two-step modeling approach proposed for the D2PMM is to be used in real applications.

A Preliminary Simulation Study

Pseudo-populations of $n = 2{,}000$ and $n = 500$ simulated responses were generated according to the D2PMM in Equation 2 for four test lengths: $n = 10$, $n = 20$, $n = 40$, and $n = 60$ by using MATLAB (1999) programs written by the author. In all cases, the item locations were uniformly distributed between -3 and $+3$, and the item discriminations ranged from 0.2 to 1.5 with a mean of 0.9. First, the simulated responses were calibrated according to the standard 2P normal-ogive model by using the program NOHARM 4 (Fraser & McDonald, 2012). Goodness of fit was assessed with two statistics: the root mean square residual (RMSR) and the GFI (goodness-of-fit index; see McDonald, 1999). The recovery of the true parameter values was assessed by computing the product–moment correlation and the mean square error (MSE) between the true and the estimated item parameters. The results are in the upper panel of Table 2.

Results in Panel (a) of Table 2 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 (specially the item locations). Given the above results on the closeness of the first- and second-order proportions of endorsement implied by the D2PMM and the 2P normal-ogive model, this result is expected and supports the appropriateness of using the last model for calibration purposes. Furthermore, the general trend found in the present results is that estimation accuracy and goodness of model-data fit increase with test length and sample size, which is reasonable.

Next, for the case of $n = 2{,}000$, EAP estimates of θ and γ were obtained on the basis of the item calibration estimates by using a MATLAB program and with the following specifications: the prior for θ was $N(0, 1)$, the prior for γ was $\chi(5, 5)$, and the number of quadrature points was 60 (see the online appendix).

As discussed above, a key practical issue is whether a minimal precision of the γ estimates can be attained in realistic situations so that θ estimates can be better than those provided by the standard 2PM. Results in Panel (b) of Table 2 are intended to assess this point. The first row of the table contains the marginal reliability estimates of θ and γ obtained by using the averages of the terms in Equation 8. Next to these estimates and within parentheses, there is the predicted reliability obtained from Table 1 by using the $\theta = 0$ and $\gamma = 1$ mean values. The second row shows the product–moment correlations between the true θ values that had been used for generating the data, and the EAPθ estimates derived from (a) the 2PM and (b) the D2PMM.

Table 2. Summary of Results of the Simulation Study.

Panel (a): Calibration Results.

	n = 10	n = 20	n = 40	n = 60
n = 2,000				
a-parameter	$r_{\hat{a},a} = .910$ $MSE_{\hat{a},a} = 0.04$	$r_{\hat{a},a} = .978$ $MSE_{\hat{a},a} = 0.02$	$r_{\hat{a},a} = .981$ $MSE_{\hat{a},a} = 0.01$	$r_{\hat{a},a} = .982$ $MSE_{\hat{a},a} = 0.01$
b-parameter	$r_{\hat{b},b} = .990$ $MSE_{\hat{b},b} = 0.01$	$r_{\hat{b},b} = .995$ $MSE_{\hat{b},b} = 0.01$	$r_{\hat{b},b} = .996$ $MSE_{\hat{b},b} = 0.01$	$r_{\hat{b},b} = .996$ $MSE_{\hat{b},b} = 0.01$
Goodness of fit	RMSR = 0.005; GFI = 0.99	RMSR = 0.005; GFI = 0.99	RMSR = 0.005; GFI = 0.99	RMSR = 0.004; GFI = 0.99
n = 500				
a-parameter	$r_{\hat{a},a} = .901$ $MSE_{\hat{a},a} = 0.04$	$r_{\hat{a},a} = .940$ $MSE_{\hat{a},a} = 0.03$	$r_{\hat{a},a} = .961$ $MSE_{\hat{a},a} = 0.02$	$r_{\hat{a},a} = .972$ $MSE_{\hat{a},a} = 0.02$
b-parameter	$r_{\hat{b},b} = .990$ $MSE_{\hat{b},b} = 0.03$	$r_{\hat{b},b} = .991$ $MSE_{\hat{b},b} = 0.02$	$r_{\hat{b},b} = .994$ $MSE_{\hat{b},b} = 0.02$	$r_{\hat{b},b} = .995$ $MSE_{\hat{b},b} = 0.02$
Goodness of fit	RMSR = 0.007; GFI = 0.97	RMSR = 0.006; GFI = 0.98	RMSR = 0.006; GFI = 0.98	RMSR = 0.006; GFI = 0.98

Panel (b): Scoring Results.

	n = 10	n = 20	n = 40	n = 60
Marginal reliability estimates	$r_{\theta,\theta} = 0.73\,(0.72)$ $r_{\gamma,\gamma} = 0.19\,(0.17)$	$r_{\theta,\theta} = 0.88\,(0.84)$ $r_{\gamma,\gamma} = 0.46\,(0.36)$	$r_{\theta,\theta} = 0.93(0.92)$ $r_{\gamma,\gamma} = 0.60(0.54)$	$r_{\theta,\theta} = 0.96(0.94)$ $r_{\gamma,\gamma} = 0.70(0.65)$
Correlation with true θ values	D2PMM $- r_{\hat{\theta},\theta} = .84$ 2PM $- r_{\hat{\theta},\theta} = .83$	D2PMM $- r_{\hat{\theta},\theta} = .91$ 2PM $- r_{\hat{\theta},\theta} = .89$	D2PMM $- r_{\hat{\theta},\theta} = .95$ 2PM $- r_{\hat{\theta},\theta} = .93$	D2PMM $- r_{\hat{\theta},\theta} = .97$ 2PM $- r_{\hat{\theta},\theta} = .95$
$MSE_{\hat{\theta},\theta}$(high γ)	0.52	0.42	0.30	0.27
$MSE_{\hat{\theta},\theta}$(low γ)	0.56	0.48	0.37	0.34

Note. RMSR = root mean square residual; GFI = goodness-of-fit index; D2PMM = dual two-parameter multiplicative model; 2PM = two-parameter model.

Finally, two sub-groups of *simulees* were formed. The upper group contained the top 27% of simulees with the highest discriminations, and the lower group contained the bottom 27% with the smallest discriminations (see Cureton, 1957). The third row in Panel (b) of Table 2 contains the MSEs between the true and the estimated θs for the upper and lower groups, respectively.

The results in Panel (b) of Table 2 can be summarized as follows. First, as far as the marginal reliabilities are concerned, the θ scores show acceptable values as from 20 items whereas, as expected, the reliability of the γ estimates requires a relatively long test to arrive at acceptable values. Even with the modest reliability of the γ estimates, however, the second and third rows in the table show that the D2PMM-based θ estimates are closer to the true θ values than the estimates based on the standard 2PM in all conditions. Admittedly, the differences are small, and further intensive research is needed if more solid conclusions are to be drawn, but the results suggest that the D2PMM can lead to improvements in trait estimation in practical applications.

Illustrative Example

The real-data study in this section uses a data set that was partly assessed in Ferrando (2007) and which consists of a test made up of 100 extraversion items taken from the various Eysenck questionnaires Maudsley Personality Inventory (MPI), Eysenck Personality Inventory Form-A (EPI-A), Eysenck Personality Inventory Form-B (EPI-B), Eysenck Personality Questionnaire Revised (EPQ-R), and Eysenck Personality Profiler (EPP) (see, for example, Miles & Hempel, 2003) administered to a sample of 531 undergraduate students. The estimated reliability of the number-correct scores (alpha) was $r_{xx} = .92$.

Items were first calibrated according to the 2P normal-ogive model by using NOHARM 4. The fit at this stage was considered to be acceptable: RMSR = 0.013 and GFI = 0.90. The average of the item discrimination estimates was 0.60, and the item locations ranged from −4 to 4 with a mean of −0.50. Overall then, the items were moderately discriminating and widely spread on the trait continuum.

EAP person estimates were next obtained based on the 2P normal-ogive model and the D2PMM. In both cases, the prior for θ was standard normal, whereas in the D2PMM, the prior for γ was χ(5, 5). Next, on the basis of these estimates, overall appropriateness and individual appropriateness (i.e., person fit) were assessed. The value of the LR Q index was 1,033.70 (with 531 degrees of freedom as reference), the average value of Λi was 0.60, and, finally, the γ estimates ranged from 0.28 to 1.68. Taken together, these results suggest that individual differences in discrimination are non-negligible in this case, so the D2PMM is more appropriate than the standard 2PM. As for person fit, only one respondent had a negative c-r_{perbis} value. So, the estimates of practically all the respondents can be validly interpreted.

To illustrate interpretation, the results of two respondents are now discussed. The θ and γ estimates for Respondent 103 were 0.16 and 1.67, respectively, whereas the corresponding estimates for Respondent 486 were 0.16 and 0.43. So, both respondents had the same trait point estimate. However, the person discrimination estimate of respondent 103 was considerably higher, which indicates substantially more deterministic responding. In accordance with this result, the $PSD(\theta)$ estimates were 0.17 for Respondent 103 and 0.42 for Respondent 486, and the corresponding reliabilities were 0.97 and 0.82, respectively. Finally, the resulting confidence bands or 68% confidence intervals (i.e., $\hat{\theta} \pm PSD(\hat{\theta})$) were [−.01, .33] for Respondent 103 and [−.26, .58] for Respondent 486. Overall, these results suggest that Respondent 103 answered the test in a more consistent way, so her trait estimate is more accurate and better reflects the standing of this person in the trait continuum.

The rest of this section illustrates two important topics discussed in the article: (a) the potential role of γ as a moderator and (b) the additional information provided by γ. As for the former,

a split-half schema was used because no external variable was available for θ. The 100 items were arranged in order of magnitude of the location values, split into odd- and even-numbered items, and the split-half correlation was computed for the θ estimates. For the entire group of 531 respondents, the correlation was $r = .88$ (the estimated reliability of the EAP estimates based on all 100 items was $r_{xx} = .94$). Then, two sub-groups were formed again using Cureton's 27% rule. The upper group contained the 143 respondents with the highest discriminations, and the lower group contained the 147 with the lowest. For the upper group, the split-half correlation was $r = .92$. For the lower group, it was $r = .82$. So, the results are in the expected direction and suggest that (a) the trait estimates of the most discriminating respondents are more accurate than those of the least discriminating respondents and (b) the individual estimates of discrimination are useful in moderated prediction.

In this data set, a 23-item Conscientiousness (C) Scale taken from the International Personality Item Pool (Goldberg, 1999) had also been administered to the 531 respondents. As mentioned above, C is expected to be (positively) related to person discrimination, and the results agreed with this expectation. The uncorrected product–moment correlation between the γ_i EAP estimates and the C Scale scores was $r = .31$ ($p < .00001$) and the bootstrap-based 90% confidence interval was [.245, .376]. The corresponding disattenuated estimate obtained by using the marginal reliability of the γ estimates and the estimated reliability of the C scores ($r_{xx} = .86$) was *r-dis* = .41.

Discussion

The starting point of this article was that, in personality measurement, an IRT model with item and person discriminations is more realistic and possibly more appropriate than the standard models in common use. If it is, a model of this type has potential advantages regarding (a) trait estimation, (b) additional information about the respondent, and (c) validity assessment. Intensive research based on real data is needed to see whether the potentially greater appropriateness and advantages are realized in practice. If this research is to be undertaken, two conditions should ideally be met. First, the additional complexity of the dual model with respect to the standard models should be minimal. Second, the dual model should be as easy to estimate as a conventional IRT model.

The present proposal has attempted to get as close as possible to the ''ideal'' conditions above. In particular, the calibration process that is proposed for the D2PMM is the same as that for the standard 2PM. And, as for the scoring, a robust and non-iterative procedure has been proposed so that person estimates are expected to be reasonable in all cases. The results of the simulation study suggest that the approach works as expected, and the real example illustrates one case in which the ''a priori'' advantages of the proposal are realized. However, many more studies on both types are needed.

The simple proposal made here can undoubtedly be improved in many ways. As mentioned above, full-information procedures with good statistical properties could be adapted for item calibration. Also, procedures for improving the individual estimates such as updating the priors could be considered. In any case, these improvements would also require simple, user-friendly, and non-commercial software to be developed so that the D2PMM could be routinely used in applied personality research.

Even though the present proposal can be improved to some extent, its basic requirements should be taken into account. The D2PMM is expected to work well in the case of relatively long tests, made up of relatively discriminating items, and with a wide spread of item locations.

As a rule of thumb, the present results suggest that even in favorable conditions useful and reasonable estimates cannot be expected with fewer than 20 items.

The required conditions discussed above might be difficult to satisfy in practice, especially because there seems to be a trend in personality measurement to use shorter and shorter tests (e.g., Emons, Sijtsma, & Meijer, 2007). There are, however, two major problems with this trend. First, as noted by Emons et al. (2007), personality tests of fewer than 20 items are generally too short even to allow trait levels to be accurately estimated. Second, the use of only a few items generally leads to a narrow-bandwidth test with a limited representation of the construct, which, in turn, is expected to lead to poor validity results. The proposal is potentially more problematic to administer in the case of computerized adaptive tests (CAT), because apart from the problem of length, the spread of item locations is modest at best (van Krimpen-Stoop & Meijer, 2002).

Overall, the problems discussed so far are parallel to those encountered for the accurate detection of person misfit (Ferrando, 2004). So, some of the solutions proposed in the person-fit framework might also be considered here. First, for those measures that have a dominant general dimension and which can be fitted, for example, by higher order or bifactor solutions, a multidimensional extension of the general D2PMM that uses information from all the test items could be considered. Second, and especially within the CAT environment, multiple administrations of subscales or pieces of information based on different item sets with different locations (e.g., easy-medium-difficult) could also be envisaged (Emons et al., 2007; van Krimpen-Stoop & Meijer, 2002).

In theory, the problems created by few items could also be mitigated by using graded response items instead of binary items, because this format is expected to increase the amount of information and so decrease measurement error. Furthermore, this type of item is indeed very common in personality and attitude measurement. For (approximately) continuous items, Ferrando (2013) proposed a dual model that is closely related to the D2PMM. However, no models of this type appear to exist for the graded response case, and it would be highly advisable to develop them. Although this need is clear, however, two caveats are in order. First, in practice, and for the problems considered here, the gains of using more continuous formats might be small in most cases (Emons et al., 2007). Second, binary items still have certain advantages in personality (e.g., Guilford, 1959).

As mentioned above, the D2PMM would not be appropriate for items that operate using a proximity mechanism. Rather, for this type of item, ideal-point models (e.g., Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009) should be more appropriate. In principle, the inclusion of an additional parameter that reflects the degree of discrimination or accuracy of the respondent has the same meaningfulness and relevance in a dominance model as in an ideal-point model. So, to develop ''dual'' versions of existing ideal-point models seems to be an issue of interest for future research.

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 research was supported by a grant from the Spanish Ministry of Economy and Competitivity (PSI2014-52884-P).

References

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika*, *67*, 49-78.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods*, *18*, 252-275.

Cudeck, R., Harring, J. R., & du Toit, S. H. (2009). Marginal maximum likelihood estimation of a latent variable model with interaction. *Journal of Educational and Behavioral Statistics*, *34*, 131-144.

Culpepper, S. A. (2010). Studying individual differences in predictability with gamma regression and nonlinear multilevel models. *Multivariate Behavioral Research*, *45*, 153-185.

Cureton, E. E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika*, *22*, 293-296.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105-113.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105-120.

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. (2013). A general linear framework for modeling continuous responses with error in persons and items. *Methodology*, *9*, 150-161.

Fiske, D. W. (1968). Items and persons: Formal duals and psychological differences. *Multivariate Behavioral Research*, *3*, 393-401.

Fraser, C., & McDonald, R. P. (2012). *NOHARM 4: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. http://www.noharm.software.informer.com

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, The Netherlands: Tilburg University Press

Guilford, J. P. (1959). *Personality*. New York, NY: McGraw-Hill.

Kelderman, H., & Molenaar, P. C. M. (2007). The effect of individual differences in factor loadings on the standard factor model. *Multivariate Behavioral Research*, *42*, 435-456.

Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement*, *33*, 545-563.

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.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, *1*, 477-482.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement*, *4*, 1-7.

MATLAB. (1999). MATLAB 5.3 Release 11.*1*. Natick, MA: The Math Works.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York, NY: Springer.

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

Miles, J. N. V., & Hempel, S. (2003). The Eysenck Personality Scales: The Eysenck Personality Questionnaire Revised (EPQ-R) and the Eysenck Personality Profiler (EPP). In M. Hersen (Ed.), *Comprehensive handbook of psychological assessment (CHOPA), Vol. 2: Personality and psychopathology assessment* (pp. 99-107). Hoboken, NJ: John Wiley.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391-411.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, *35*, 543-568.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217-226.

Rizopoulos, D., & Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, *61*, 415-438.

Stark, S., Chernyshenko, O., Drasgow, F., & Williams, B. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25-39.

Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement*, *11*, 355-370.

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

Tay, L., Drasgow, F., Rounds, J., & Williams, B. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology*, *94*, 1287-1304.

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.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*, 164-180.

Voyce, C. D., & Jackson, D. N. (1977). An evaluation of threshold theory for personality assessment. *Educational and Psychological Measurement*, *37*, 383-408.