Psicothema

Psicothema 2016, Vol. 28, No. 1, 76-82
doi: 10.7334/psicothema2015.204

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2016 Psicothema
www.psicothema.com

Traditional scores versus IRT estimates on forced-choice tests based on a dominance model

Pedro M. Hontangas[1], Iwin Leenen[2], Jimmy de la Torre[3], Vicente Ponsoda[4], Daniel Morillo[4] and Francisco J. Abad[4]

[1] Universidad de Valencia, [2] Universidad Nacional Autónoma de México, [3] The State University of New Jersey (USA), and [4] Universidad Autónoma de Madrid

Abstract

Background: Forced-choice tests (FCTs) were proposed to minimize response biases associated with Likert format items. It remains unclear whether scores based on traditional methods for scoring FCTs are appropriate for between-subjects comparisons. Recently, Hontangas et al. (2015) explored the extent to which traditional scoring of FCTs relates to the true scores and IRT estimates. The authors found certain conditions under which traditional scores (TS) can be used with FCTs when the underlying IRT model was an unfolding model. In this study, we examine to what extent the results are preserved when the underlying process becomes a dominance model. **Method:** The independent variables analyzed in a simulation study are: forced-choice format, number of blocks, discrimination of items, polarity of items, variability of intra-block difficulty, range of difficulty, and correlation between dimensions. **Results:** A similar pattern of results was observed for both models; however, correlations between TS and true thetas are higher and the differences between TS and IRT estimates are less discrepant when a dominance model involved. **Conclusions:** A dominance model produces a linear relationship between TS and true scores, and the subjects with extreme thetas are better measured.

Keywords: forced-choice, dominance model, traditional scores, EAP.

Resumen

Puntuaciones tradicionales y estimaciones TRI en tests de elección forzosa con un modelo de dominancia. **Antecedentes:** los tests de elección forzosa (TEFs) fueron propuestos para reducir los sesgos de respuesta de ítems tipo Likert. Se cuestiona que los métodos de puntuación tradicional (PT) empleados permitan hacer comparaciones entre-sujetos. Recientemente, Hontangas et al. (2015) exploraron cómo las PTs obtenidas con diferentes TEFs se relacionan con sus puntuaciones verdaderas y estimaciones TRI, mostrando las condiciones para ser utilizadas cuando el modelo subyacente es un modelo de unfolding. El objetivo del trabajo actual es comprobar si el patrón de resultados se mantiene con un modelo de dominancia. **Método:** las variables independientes del estudio de simulación fueron: formato de elección forzosa, número de bloques, discriminación de los ítems, polaridad de los ítems, variabilidad de la dificultad intrabloque, rango de dificultad del test y correlación entre dimensiones. **Resultados:** un patrón similar de resultados fue obtenido en ambos modelos, pero en el modelo de dominancia las correlaciones entre PTs y puntuaciones verdaderas son más altas y las diferencias entre PTs y estimaciones TRI se reducen. **Conclusiones:** un modelo de dominancia produce una relación lineal entre PTs y puntuaciones verdaderas, y los sujetos con puntuaciones extremas son medidos mejor.

Palabras clave: elección forzada, modelo dominancia, puntuación tradicional, EAP.

Forced-choice tests (FCTs) have a long tradition in psychology, and were proposed to address response bias in Likert-type items (Christiansen, Burns, & Montgomery, 2005; McCloy, Heggestad, & Reeve, 2005). FCTs are generally composed by blocks of items with different items measuring different dimensions (Multidimensional, MFCTs) or the same dimension (Unidimensional, UFCTs). The most common formats are called PICK, MOLE and RANK (Hontangas et al., 2015). The PICK format instructs respondents to choose the item in the block that is most descriptive of them; in the MOLE format, they choose the most as well as the least descriptive item; and when the RANK format is used, respondents rank all the items from most to least descriptive (for examples, see Hontangas et al., 2015; and for a condensed literature review on FCTs, see van Eijnatten, van der Ark, & Holloway, 2015).

Empirical evidence shows that MFCTs, to some extent, can control response biases (e.g., Cheung & Chan, 2002; Saville & Willson, 1991), increase criterion validity (Bartram, 2007; Brown & Maydeu-Olivares, 2013), and avoid or reduce faking (Christiansen et al., 2005; Heggestad, Morrison, Reeve, & McCloy, 2006; Hirsh & Peterson, 2008; Jackson, Wroblewski, & Ashton, 2000). However, MFCTs have been criticized because traditional scoring methods yield ipsative or partially ipsative scores and typically result in incorrect estimates of reliability, scale intercorrelations, and factor loadings, and in unwarranted (normative) interpretations

Received: July 25, 2015 • Accepted: October 29, 2015
Corresponding author: Pedro M. Hontangas
Facultad de Psicología
Universidad de Valencia
46011 Valencia (Spain)
e-mail: hontanga@uv.es

of the obtained scores (Brown & Maydeu-Olivares, 2013; Closs, 1996; Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988; Meade, 2004). With respect to the latter, ipsative scores are considered appropriate for intra-individual comparisons only (i.e., they provide information about an individual's relative standings across multiple dimensions, and could be useful for counseling purposes), but they are not appropriate for inter-individual or normative comparisons, and, as a consequence, should not be used for personnel selection (Closs, 1996; Johnson et al., 1988). However, some authors report significant correlations between ipsative and normative scores (Matthews & Oddy, 1997; Saville & Wilson, 1991) and between ipsative scores and external criterion variables (Christiansen et al., 2005; Jackson et al., 2000), which may be considered an argument in favor of the normative interpretation of ipsative scores (Baron, 1996; see also, Clemans, 1966, for a theoretical analysis of the properties of ipsative scores).

To gain insight into how normative information can be extracted from MFCTs, two approaches have been followed. In the first approach, new models within the item response theory (IRT) framework have been proposed, such as the multi-unidimensional pairwise-preference model (MUPP; Stark, Chernyshenko, & Drasgow, 2005), the McCloy, Heggestad & Reeve (2005) model, the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011), or the MUPP-2PL model (Morillo et al., 2015). The second approach relies on simulation studies to examine the correspondence between empirical scores, obtained using different FC formats and/or scoring procedures, and true scores. Three such studies have been reported: Matthews and Oddy (1997) and Saville and Willson (1991) only consider a particular FC format and employ *ad hoc* procedures to generate true scores and empirical scores; Hontangas et al. (2015), on the other hand, systematically examined this correspondence using a large number of conditions—block format, test length, item properties (discrimination, location, polarity, intra-block variability), and correlations between traits—relying upon one of the above mentioned IRT models for the data generation process. The authors examined how traditional scores (TS; i.e., ipsative scores) using the MFCT formats most used in applied context (i.e., PAIR for two-item blocks; PICK, MOLE and RANK for four-item blocks) relate to the true scores (θ) and the IRT estimates (EAP). These estimates are considered as an upper limit to show the efficiency of recovery lost when TS are used or to examine to what extent they order the respondents as IRT estimates do. Although the overall mean correlations with the true scores were moderate (r_{TS} = .62, r_{EAP} = .86), satisfactory results were obtained in particular conditions (e.g., with independent traits, large tests, high discrimination, different polarity, and MOLE format, r_{TS} = .87 and r_{EAP} = .98 was found). The study's main conclusions were: a) the RANK format works best compared to other formats, with the MOLE format coming in as a close second; b) in practice, the MOLE format might be more viable because it is less demanding for the respondents; c) a MFCT should have a large number of blocks and highly discriminating items that adequately cover the levels of the traits considered; and d) a MFCT works better with balanced blocks (with positive and negative items), and with independent rather than correlated traits.

Hontangas et al. (2015) employed the original MUPP model by Stark et al. (2005) and its extension to more than two-items blocks (de la Torre, Ponsoda, Leenen, & Hontangas, 2011) for the data generation process. The MUPP model is an unfolding model (UM) as it incorporates the generalized graded unfolding model (GGUM, Roberts, Donoghue, & Laughlin, 2000), which implies an ideal-point process for the latent response of an individual to each item in the block (i.e., the probability that an individual endorses an item increases with the proximity between the item's position and the individual's ideal point on the latent dimension measured by the item). UMs contrast with dominance models (DMs), where the response process implies a monotone relation between the latent trait and the probability of endorsement. Several studies with simulated (Liao & Mead, 2009; Tay, Ali, Drasgow, & Williams, 2011) and empirical (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Chernyshenko, Stark, Drasgow, & Roberts, 2007; Sherbaum, Finlinson, Barden, & Tamanini, 2006; Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009; Weekers & Meier, 2008) data provide evidence that in noncognitive domains, like personality, attitudes, or interests, an UM (and specially the GGUM model) is as good as or better than a DM. An UM is considered better because it is more flexible as it can be equivalent to DM when items have location parameters at the end of the continuum (Stark, Chernychenko, & Drasgow, 2005; Drasgow, Chernyshenko, & Stark, 2010). However, other, more recent studies shown that this superiority is not generally the case in practice. Huang and Mead (2014) found that, compared to dominant items, scales composed entirely of ideal-point items had substantially inferior psychometric properties, including lower score reliabilities and lower correlations with external criteria. Furthermore, Carvalho, Filho, Pessotto and Bortolotti (2014) show that, when items coincide with the extremes of the latent trait continuum (e.g., when they measure pathological aspects of personality), an UM does not offer advantages over a DM. The same idea was put forward by Brown and Maydeu-Olivares (2010) and Oswald and Shell (2010), who argued that the added value of UMs comes from items situated in the middle of the trait continuum, but these items tend to be complex or multidimensional, they are difficult to write, and the exact meaning of the response is unclear. Moreover, Brown and Maydeu-Olivares (2010) argue that an UM is not invariant to reverse scoring and the estimation may not be as accurate as a DM, and Oswald and Shell (2010) consider that parsimonious models are preferred over complex ones, that is, we should stick to the simpler models unless clear evidence exists of the advantages of complex models. Additional arguments on the controversy 'dominance versus unfolding' can be found in the article by Drasgow et al. (2010) and the subsequent commentaries in the same issue.

Most operational MFCTs have been elaborated with dominance items (Brown & Bartram, 2009; Brown & Maydeu-Olivares, 2010) as they are based on classical test theory, where item selection criteria advise to discard nonmonotonic (ideal point process) items because they are characterized by low or negative item-total correlations or weak factor loadings. However, the properties of TS obtained from MFCTs with an underlying dominance process have not yet been studied. The aim of this paper is to examine to what extent the pattern of results obtained using an UM (Hontangas et al., 2015) is preserved when the underlying model is a DM. Thus, the study will provide a more realistic description of the effects of TS as obtained from current MFCTs, given that, as argued above, most of these tests are built with items that adhere to a dominance response process rather than to an ideal-point response process.

Method

In general terms, this study follows the same procedure and conditions described in Hontangas et al. (2015), except that a DM rather than an UM is used for the data generation.

IRT models

For the PAIR format, the study used the MUPP-2PL model (Morillo et al., 2015). It was chosen because it uses discrimination and location parameters like GGUM model. For the PICK, MOLE and RANK formats, an extension of the latter model to four-item blocks is employed (along the lines suggested by de la Torre et al., 2011). Note that the MUPP-2PL is similar to the original MUPP (Stark et al., 2005), except that the GGUM (an UM) is replaced by Birnbaum's (1968) two parameter logistic model (2PLM, a DM).

Simulation study

Design. The following factors were varied: (a) block format: PAIR, PICK, MOLE and RANK; (b) test length: 18 and 36 blocks; (c) item discrimination (*a* parameter): low and high; (d) item polarity (sign of the *a* parameters): all positive, half-and-half, or mixed blocks; (e) variability of difficulty (standard deviation of the *b* parameters within a block): low (.2) and high (.8); (f) range of the *b* parameters in the tests: –2 to 2 and –3 to 3; and (g) true θ correlation ϱ: 0 and .5. The number of dimensions was fixed at four; and responses of 5,000 respondents were generated so as to improve the stability of the results.

For the PICK, MOLE and RANK formats, the four dimensions were measured in each block; for the PAIR format, each dimension was measured in 9 (short test) or 18 blocks (long test), and 3 or 6 blocks measured each of the six possible pairs of dimensions. The values of the *a* parameters were sampled from a continuous uniform distribution, U[.75, 1.25] for low discrimination, and from U[1.75, 2.25] for high discrimination items (since both models are not equivalent, these values should be interpreted as minor and major discrimination). Item polarity was defined by the sign of the *a* parameters: For the half-and-half condition, one-half of the blocks had positive *a* parameters for all items and the other half had negative *a*'s. For the mixed condition, 1/3 of the blocks were positive, 1/3 were negative, and the remaining 1/3 contained both positive and negative items. The positive and negative *a* parameters are balanced in the last two polarity conditions to ensure that each dimension was measured by approximately the same number of positive and negative items. The difficulty parameters were drawn either from U[–2, 2] (which coincides with the range for the delta parameters used in the UM by Hontangas et al., 2015), or from U[–3, 3] (which corresponds to frequently used ranges for the difficulty parameter in the 2PLM, see Harris, 1989). Finally, the θ parameters were generated from a multivariate normal distribution with a mean vector of zero, variances of one, and a Pearson correlation of ϱ between any two dimensions.

Scoring procedures. IRT scores were obtained using the expected a posteriori (EAP) estimation procedure (Bock & Mislevy, 1982). For TS involving the PAIR and PICK formats, the selected dimension received 1 point when the item was positive and –1 point when it was negative; the nonselected dimensions received 0 points. For the MOLE format, the dimension of the *most* preferred item received 1 point when it was positive (–1, if negative), and that of the *least* preferred –1 point when it was positive (1, if negative); the nonselected items provided 0 points to their dimensions. For the RANK format, if the item was positive, the most preferred dimension received 4 points; the second most preferred dimension, 3 points;… and the least preferred, 1 point; if the item was negative, the most preferred received 1 point; the second most preferred, 2 points; ... and the least preferred, 4 points. In all formats, the test score on each dimension was the sum across blocks of the corresponding dimension scores.

Measures of goodness of parameter recovery. Three measures (bias, RMSE and the Pearson correlation between true and test scores) were computed to evaluate the goodness of recovery of the latent trait parameter. All the measures were computed for the IRT estimates, but only the Pearson correlation was computed for TS.

Results

First, we note that the range of the difficulty parameters (–2 to 2 vs. –3 to 3) was not a relevant factor: the difference in all measures of goodness of parameter recovery for both cases are negligible and this factor nor as a main effect nor its interaction with other factors has a significant effect. In order to facilitate the comparison with the results from the UM (Hontangas et al. 2015; see below), we will present results only for the range –2 to 2 (which was also the range used for the difficulty parameters in the UM). Table 1 shows the mean correlations of TS and EAP estimates with the true scores for each factor and some interactions; it is clear from this table that EAP performed better than TS in all conditions.

Although the RANK format yielded the best results (r_{TS} = .74; r_{EAP} = .85, RMSE = .36), they were very similar to those for the MOLE format, not only overall (r_{TS} = .73; r_{EAP} = .85, RMSE = .36) but in any of the conditions (with the difference between RANK and MOLE correlations always being less than .023 for TS and .004 for EAP). Both showed better results than the PICK (r_{TS} = .69; r_{EAP} = .77, RMSE = .53) and PAIR (r_{TS} = .68; r_{EAP} = .73, RMSE = .62) formats, with the latter two again being similar (the difference between the corresponding correlations never exceeds .038 for TS and .055 for EAP). The bias of the EAP estimates (i.e., the mean algebraic difference with the true thetas) was found to be very small (PAIR = –.0016, PICK = –.0010, MOLE = –.0009, and RANK = –.0011) and unrelated with any of the conditions. However, a slight conditional bias for extreme negative and positive thetas was observed (see Figure 1a). However, this effect is well known and generally expected in the EAP estimation method (Kim & Nicewander, 1993); note, for example, that such bias was also present in the UM (see Figure 1b; Hontangas et al., 2015).

Considering the remaining conditions, the same pattern of results was also observed for both models. Summarizing the findings for the remaining conditions, we found that both TS and EAP estimates were better when: a) the test has more blocks (36 blocks: r_{TS} = .73, r_{EAP} = .83; 18 blocks: r_{TS} = .70, r_{EAP} = .77), b) the items had higher discrimination (high: r_{TS} = .73, r_{EAP} = .81; low: r_{TS} = .70, r_{EAP} = .79), c) the polarity of items was mixed (mixed: r_{TS} = .80, r_{EAP} = .91; half-and-half: r_{TS} = .68, r_{EAP} = .75; positive: r_{TS} = .67, r_{EAP} = .74), and d) the dimensions were independent (ϱ = 0: r_{TS} = .72, r_{EAP} = .88; ϱ = .5: r_{TS} = .52, r_{EAP} = .85). The largest differences, especially for TS and PAIR and PICK models, were found among the levels of *polarity* and *trait correlations*. The effect of the intra-block variability among difficulty parameters is small (high: r_{TS} = .72, r_{EAP} = .79; low: r_{TS} = .71, r_{EAP} = .80).

Comparing the results for the DM and the UM

The main finding of a comparison between the results just presented for the DM (current study) and those obtained within the context of an UM (Hontangas et al., 2015) is that in the DM the correlation between the TS and the true thetas substantially improved, while the correlation between the true thetas and the EAP estimates slightly decreased (Table 2). In particular, in the UM study the overall mean correlation of the true scores with the TS and EAP estimates were .62 and .86, respectively, whereas these correlations changed to .71 and .80 in the DM. That is, an increase of .09 in the correlation for the TS and a decrease of .06 for in the UM. For the different factors in the simulation study, we observe that the improvement for the TS is more pronounced for PAIR and PICK formats, lower number of blocks, lower discrimination, and positive polarity. The largest effect resulted from combinations of blocks with positive items and low variability, and the differences due variability of the item difficulty were quite similar (Table 2).

On the other hand, the superiority of EAP estimates versus TS significantly decreased, i.e., the differences of correlations (r_{EAP} - r_{TS}) were lower in the DM (overall: UM = .24, DM = .09; PAIR: UM = .20, DM = .05; PICK: UM = .26, DM = .08; MOLE: UM = .26, DM = .11; RANK: UM = .26, DM = .10).

Looking at the results from the DM study and the UM study from a different perspective, it is interesting to note that correlations between TS and EAP estimates are substantially higher under the DM than under the UM (overall: r_{DM} = .89, r_{UM} = .68; PAIR: r_{DM} = .93, r_{UM} = .70; PICK: r_{DM} = .90, r_{UM} = .66; MOLE: r_{DM} = .86, r_{UM} = .68; RANK: r_{DM} = .87, r_{UM} = .69), which indicates that the TS and EAP estimates behave more similarly under the DM. By way of example, Figure 1e and 1f show a scatter plot that illustrates the relation between both estimates.

Apart from the differences between the DM and UM, importantly, the *pattern* of results turns out to be largely the same for both models: (a) EAP correlate higher with the true scores than TS do; (b) the RANK and MOLE format yielded quite similar results, as the PICK and PAIR formats do, and the results for the former were clearly better than for the latter; (c) MFCTs with more blocks, consisting of a mix of positive and negative items that have high discrimination parameters result in

Table 1
Mean correlations between True Thetas and Traditional Scores or EAP estimates (range of difficulty: from -2 to 2)

	Traditional scores				**EAP**			
	PAIR	**PICK**	**MOLE**	**RANK**	**PAIR**	**PICK**	**MOLE**	**RANK**
Overall								
UM*	.571	.575	.662	.669	.768	.831	.924	.927
DM	.684	.692	.733	.744	.730	.771	.846	.847
Blocks								
18	.657	.673	.722	.734	.696	.741	.823	.825
36	.712	.711	.744	.754	.763	.802	.869	.869
***a* parameters**								
(.75, 1.25)	.655	.677	.725	.736	.706	.756	.844	.846
(1.75, 2.25)	.714	.708	.741	.752	.753	.787	.848	.848
Blocks x *as*								
18 (.75, 1.25)	.619	.651	.710	.721	.664	.717	.814	.818
(1.75, 2.25)	.694	.696	.735	.746	.728	.765	.832	.832
36 (.75, 1.25)	.690	.703	.741	.750	.748	.794	.873	.874
(1.75, 2.25)	.733	.719	.747	.757	.778	.809	.864	.864
Polarity								
All positive	.636	.635	.697	.702	.671	.706	.795	.797
Half-and-half	.636	.670	.697	.702	.669	.722	.795	.797
Mixed	.781	.771	.806	.827	.850	.886	.947	.947
***b* parameters: Variability**								
.2	.684	.687	.728	.738	.736	.775	.849	.851
.8	.684	.697	.738	.749	.724	.768	.842	.843
Polarity x Variability								
All positive .2	.641	.629	.696	.700	.679	.710	.802	.804
.8	.631	.642	.697	.703	.662	.702	.788	.789
Half-and-half .2	.641	.671	.696	.700	.677	.727	.798	.801
.8	.631	.669	.699	.705	.661	.716	.792	.793
Mixed .2	.771	.762	.793	.814	.851	.886	.946	.947
.8	.791	.780	.818	.841	.848	.887	.947	.948
Θs correlations								
0	.784	.797	.840	.848	.799	.832	.889	.890
.5	.584	.587	.626	.639	.660	.711	.802	.804

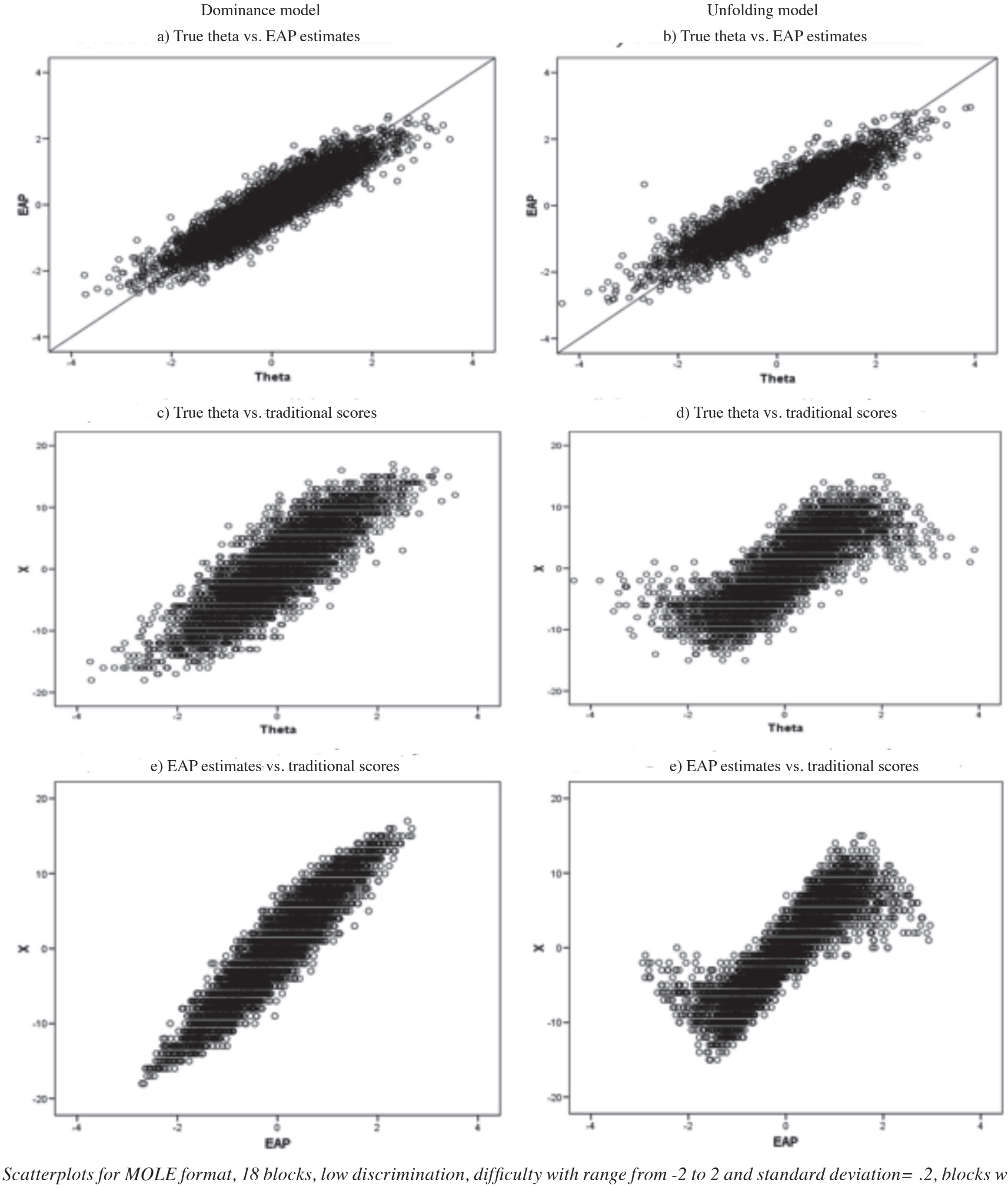
estimates that better predict the true scores; and, finally, (d) the effect of the intra-block variability among the items' difficulties was negligible.

Discussion

In this paper, we examined how the properties of TS are affected if a DM rather than an UM underlies the responses on a MFCT. This is a relevant inquiry because the response process in most operational MFCTs is likely to be closer to the one assumed by a DM.

In general, although the *same pattern* of results was observed for both models, assuming a dominance process yielded a *substantial improvement* for the TS as compared to the same estimates in the UM, that is, correlations between TS and true thetas were substantially higher and the differences between TS and IRT estimates were reduced.

Figure 1 may provide an explanation for this improvement: A DM produces a more linear relationship between TS and true scores (Figure 1c), whereas under an UM, this relationship is obviously nonlinear, especially at the extremes (Figure 1d). Obviously, this deviation from linearity is tantamount to the lower correlation


Dominance model

Unfolding model

a) True theta vs. EAP estimates

b) True theta vs. EAP estimates

c) True theta vs. traditional scores

d) True theta vs. traditional scores

e) EAP estimates vs. traditional scores

e) EAP estimates vs. traditional scores

Figure 1. Scatterplots for MOLE format, 18 blocks, low discrimination, difficulty with range from -2 to 2 and standard deviation= .2, blocks with mixed items, and independent traits

between TS and true scores, which especially translates to poor estimation of respondents with extreme thetas in the UM. (As an aside, we note that the DM and UM performed equally well with respondents who had intermediate theta's.)

The latter considerations, however, do not imply that the advantage of a DM over a UM is related to a possibly inappropriate range for the difficulty parameters (which translates to not having items appropriately adjusted to subjects with an extreme position on the latent trait) because similar results were obtained with a larger range for the difficulty parameter (–3 to 3). A more plausible explanation is that, implicitly, the TS are also based on a dominance process. For all formats, TS on each dimension are obtained by the sum across blocks of the corresponding dimension scores. As such, there is a monotonic relationship between total score and level of trait estimate (i.e., a higher score indicates a higher trait level, and a lower score implicates a lower level). As a result, because the model for generating the data was congruent with the underlying process assumed in the scoring procedure (i.e., both are dominance), results were better than when the response and scoring processes were incongruent (i.e., unfolding data analyzed with a dominance model).

One limitation of the present study is that the DM and UM cannot be directly compared, because parameters, even if they have the same name, bare different interpretations. For example, we are aware that, although both models contain an item location parameter, they actually have different meanings. For the UM, the location parameter (i.e., *delta*) indicates the trait level where the probability of endorsing the statement is *the highest*; for the DM, the location parameter (i.e., *b*) indicates the trait level where there is a *50% chance* of endorsing the item. Although we chose item parameters that made the two models as similar as possible—we used parameter values that have been usually employed with these models, and ranges that have approximately the same values—, we acknowledge that the models may still be not comparable. Future research should explore more theoretically and formally how the correspondence between the two models can be established. It would also be interesting to explore why the extreme thetas are measured poorly in the UM. Finally, although we confirmed that there are conditions where some normative information can be extracted from MFCTs, we do not advise in favor of the use of traditional scoring methods for inter-individual comparisons because they have problems related to reliability and validity that can be better addressed by using the appropriate IRT models.

Acknowledgements

The research has been funded by the Ministry of Economy and Competitivity of Spain, project PSI2012-33343.

Table 2
Differences between mean correlations of both models (DM – UM)

	Traditional scores				EAP			
	PAIR	PICK	MOLE	RANK	PAIR	PICK	MOLE	RANK
Overall	.113	.118	.071	.075	-.038	-.060	-.078	-.079
Blocks								
18	.118	.126	.076	.080	-.020	-.045	-.080	-.081
36	.109	109	.066	.070	-.057	-.075	-.077	-.077
***a* parameters**								
(.75, 1.25)	.119	.128	.079	.082	-.002	-.021	-.054	-.056
(1.75, 2.25)	.107	.108	.064	.068	-.075	-.098	-.103	-.103
Blocks x *a*s								
18 (.75, 1.25)	.123	.137	.085	.088	.016	.-.003	-.051	-.054
(1.75, 2.25)	.112	.115	.067	.071	-.057	-.087	-.108	-.109
36 (.75, 1.25)	.116	.118	.072	.076	-.019	-.040	-.056	-.057
(1.75, 2.25)	.102	.100	.060	.065	-.094	-.109	-.097	-.097
Polarity								
All positive	.146	.129	.121	.120	-.027	-.076	-.089	-.091
Half-and-half	.107	.118	.061	.063	-.100	-.116	-.139	-.140
Mixed	.086	.106	.031	.042	.011	.013	-.007	-.007
***b* parameters: Variability**								
.2	.098	.111	.076	.081	-.011	-.030	-.059	-.061
.8	.128	.124	.066	.069	-.066	-.090	-.098	-.098
Polarity x Variability								
All positive .2	.136	.132	.141	.141	.017	-.022	-.049	-.052
.8	.157	.127	.102	.100	-.071	-.131	-.129	-.129
Half-and-half .2	.099	.116	.065	.068	-.061	-.082	-.121	-.123
.8	.115	.119	.058	.058	-.139	-.150	-.157	-.157
Mixed .2	.060	.086	.022	.034	.012	.016	-.006	-.007
.8	.112	.126	.039	.049	.011	.011	-.008	-.008
Θs correlations								
0	.118	.133	.073	.077	.010	-.009	-.045	-.046
.5	.108	.103	.069	.073	-.087	-.110	-.112	-.112

References

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69,* 49-56.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*, 263-272.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Brown, A., & Bartram, D. (2009). *Doing Less but Getting More: Improving Forced-Choice Measures with IRT.* Paper presented at the 24th Annual conference of the Society for Industrial and Organizational Psychology, New Orleans.

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology*, *3*(4), 489-493.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modelling of forced-choice questionnaires. *Educational and Psychological Measurement, 71,* 460-502.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18,* 36-52.

Carvalho F., Filho, A., Pessotto, F., & Bortolotti, S. (2014). Application of the unfolding model to the aggression dimension of the Dimensional Clinical Personality Inventory (IDCP). *Revista Colombiana de Psicología, 23*, 339-349.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88-106.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562.

Cheung, M. W.-L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling*, *9*, 55-77.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*, 267-307.

Clemans, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures* (Psychometrika Monograph No. 14). Richmond, VA: Psychometric Society.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology*, *69*, 41-47.

de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2011). *Some extensions of the multi-unidimensional pairwise preference model*. Paper presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Drasgow, F. L., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*, 465-476.

Harris, D. (1989). Comparison of 1-,2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, Spring,* 34-41.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91,* 9-24.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*, 167-184.

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a "Fake-Proof" measure of the Big Five. *Journal of Research in Personality*, *42*, 1323-1333.

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*. Advance online publication, doi:10.1177/0146621615585851

Huang, J., & Mead, A. D. (2014). Effect of personality item writing on psychometric properties of Ideal-Point and Likert scales. *Psychological Assessment, 26*, 1162-1172.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, *13*, 371-388.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spouriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153-162.

Kim, J. K., & Nicewander, A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587-599.

Liao, C., & Mead, A. D. (2009, April). *Fit of ideal-point and dominance IRT models to simulated data*. Paper presented at the 24th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Matthews, G., & Oddy, K. (1997). Ipsative and normative scales in adjectival measurement of personality: Problems of bias and discrepancy. *International Journal of Selection and Assessment, 5,* 169-182.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8*, 222-248.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using Ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531-552.

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P. M., de la Torre, J., & Ponsoda, V. (2015, *submitted*). A dominance variant under the multi-unidimensional Pairwise-preference framework: Model formulation and Markov Chain Monte Carlo Estimation. *Applied Psychological Measurement*.

Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right - but so far, Likert was not wrong. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3,* 481-484.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3-32.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, *64*, 219-238.

Scherbaum, C. A., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *The Leadership Quarterly, 17*, 366-386.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184-203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (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*(1), 25-39.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models, assessing the relative model-data fit of ideal point and fominance models. *Applied Psychological Measurement, 35*, 280-295.

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.

van Eijnatten, F. M., van der Ark, L. A., & Holloway, S. S. (2015). Ipsative measurement and the analysis of organizational values: An alternative approach for data analysis. *Quality and Quantity: International Journal of Methodology*, *49*, 559-579.

Weekers, M. A., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*, 65-77.