Comparing Traditional and IRT Scoring of Forced-Choice Tests

Applied Psychological Measurement
2015, Vol. 39(8) 598–612
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615585851
apm.sagepub.com

**Pedro M. Hontangas¹, Jimmy de la Torre²,
Vicente Ponsoda³, Iwin Leenen⁴, Daniel Morillo³,
and Francisco J. Abad³**

Abstract

This article explores how traditional scores obtained from different forced-choice (FC) formats relate to their true scores and item response theory (IRT) estimates. Three FC formats are considered from a block of items, and respondents are asked to (a) pick the item that describes them most (PICK), (b) choose the two items that describe them the most and the least (MOLE), or (c) rank all the items in the order of their descriptiveness of the respondents (RANK). The multi-unidimensional pairwise-preference (MUPP) model, which is extended to more than two items per block and different FC formats, is applied to obtain the responses to each item block. Traditional and IRT (i.e., expected a posteriori) scores are computed from each data set and compared. The aim is to clarify the conditions under which simpler traditional scoring procedures for FC formats may be used in place of the more appropriate IRT estimates for the purpose of inter-individual comparisons. Six independent variables are considered: response format, number of items per block, correlation between the dimensions, item discrimination level, and sign-heterogeneity and variability of item difficulty parameters. Results show that the RANK response format outperforms the other formats for both the IRT estimates and traditional scores, although it is only slightly better than the MOLE format. The highest correlations between true and traditional scores are found when the test has a large number of blocks, dimensions assessed are independent, items have high discrimination and highly dispersed location parameters, and the test contains blocks formed by positive and negative items.

Keywords

forced choice, ipsative data, multi-unidimensional pairwise-preference, MUPP, unfolding model, GGUM, EAP, traditional scoring, personality assessment, faking

¹Universidad de Valencia, Spain

²Rutgers, The State University of New Jersey, New Brunswick, USA

³Universidad Autónoma de Madrid, Spain

⁴Universidad Nacional Autónoma de México, Mexico City, México

Corresponding Author:

Jimmy de la Torre, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901, USA.
Email: j.delatorre@rutgers.edu

Introduction

Forced-Choice (FC) Formats and Responses Biases

Likert-type format is the most popular item type for the measurement of non-cognitive constructs. It is well known that responses to this item format can be affected by various response biases such as acquiescence, extreme and central tendency responding, response-order, social desirability, and faking (Paulhus, 1991). To address the bias-related issues in Likert-type items, alternative formats, such as the FC, have been proposed (Christiansen, Burns, & Montgomery, 2005; McCloy, Heggstad, & Reeve, 2005; Travers, 1951). FC measures have a long tradition in psychology. Van Eijnatten, Van der Ark, and Holloway (2015) provided a condensed literature review on FC measures. In keeping with the terminology in recent works (e.g., Brown & Maydeu-Olivares, 2011, 2013), the term *block* will be used in this article to refer to a collection of items on which respondents can base their responses. Table 1 gives examples of blocks for the different FC formats. For these formats, respondents are instructed to (a) choose or pick the item that is most descriptive of them (PICK), (b) choose the items that are the most and least descriptive of them (MOLE, from M^Ost and L^East descriptive items), or (c) rank all the items in the order of their descriptiveness (RANK). Blocks may differ not only in the response required of the respondents but also in terms of (a) their number of items, (b) dimensionality (uni- or multidimensional, depending on the dimensions tapped by the blocks), and (c) block polarities (presence of purely positive or negative, or mixed items in a block).

A large number of papers provide empirical evidence that FC formats can better control response biases (e.g., Cheung & Chan, 2002; Edwards, 1957, 1970; Saville & Willson, 1991), and increase criterion validity (Bartram, 2007; Brown & Maydeu-Olivares, 2013). Much attention has especially been paid to the issue of faking or intentional response distortion (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006), and whether the FC format can avoid or reduce it. The empirical evidence has shown mixed results: several studies have shown that faking can be partially controlled (Christiansen et al., 2005; Hirsh & Peterson, 2008; Jackson, Wroblewski, & Ashton, 2000), although the positive effect has only been found at group-level, but not at individual-level, analyses (Heggstad, Morrison, Reeve, & McCloy, 2006).

Although previous studies indicate that the FC format assessments may provide some advantages, they have also received some criticisms because they yield ipsative or partially ipsative scores, which have problematic psychometric properties pertaining to reliability, average scale inter-correlations, factor analysis, and interpretation of scores (Brown & Maydeu-Olivares, 2013; Closs, 1996; Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988; Meade, 2004). Data obtained with an FC format are ipsative because the sum of scores across all dimensions is a constant for each participant, which means that each dimension score is dependent on other dimension scores (Hicks, 1970). A recurring debate with ipsative scores pertains to their correct interpretation. On one hand, critics contend that ipsative scores are only appropriate for intra-individual comparisons and can be used with no concern to provide information about an individual's relative standings across multiple dimensions, which can be very valuable for counseling purposes. However, they are not appropriate for inter-individual comparisons or normative interpretations, and should not be used for personnel selection (Aitchison, 2003; Closs, 1996; de Vries, 2006). On the other hand, proponents argue that normative and ipsative scores can correlate strongly, particularly when a higher number of dimensions are involved, and the inter-correlations between these dimensions are low (Clemans, 1966). As such, ipsative scores can be interpreted in a normative way (Baron, 1996). This position is supported by other studies as well, where significant correlations between ipsative scores and either several criterion

Table 1. Examples of FC Formats.

PICK

Unidimensional (Preference for Solitude Scale; Burger, 1995).

Select the statement that best describes you:

- A. I enjoy being around people B. I enjoy being by myself

Multidimensional (Part 3 of I6PF-APQ test; Schuerger, 2008).

I would prefer:

- A. To repair a broken device (washing machine, car, lawnmower or other machine)
B. To create a work of art (poem, painting, etc.) or appreciate what someone else has done

Different item polarities (Christiansen, Burns, & Montgomery, 2005).

Which of the following adjectives is most true or most descriptive of you? Are you more:

1. Practical or imaginative 2. Unkind or careless

MOLE (The Occupational Personality Questionnaire; Brown & Bartram, 2009).

Indicate which is most and least like you:

A. I hate to lose	most	least
B. I find creative ideas come easily	most	least
C. I conceal my feeling	most	least
D. I take the lead in a group	most	least

RANK (The Study of Values; Kopelman, Rovenpor, & Guan, 2003).

Viewing Leonardo da Vinci's picture, "The Last Supper," would you tend to think of it:

- A. as expressing the high spiritual aspirations and emotions
B. as one of the most priceless and irreplaceable pictures ever painted
C. in relation to Leonardo's versatility and its place in history
D. the quintessence of harmony and design

Note. FC = forced-choice; PICK = pick the item that describes them most; APQ = Adolescent Personality Questionnaire; MOLE = choose the two items that describe them the most and the least; RANK = rank all the items in the order of their descriptiveness of the respondents.

variables (Christiansen et al., 2005; Jackson et al., 2000) or normative scores (Matthews & Oddy, 1997; Saville & Willson, 1991) have been found.

To better understand whether ipsative scores can have normative utility, true scores and empirical scores, which were obtained with different FC formats and scoring procedures, were compared in a simulation study. As far as the authors know, except for a short simulation (Matthews & Oddy, 1997, appendix) of the responses to the PAL-TOPAS questionnaire (described below), this type of inquiry has never been carried out in the past.

Traditional Scoring Procedures for FC Tests

To identify the different types of FC tests in use, commercial and non-commercial tests extracted mainly from test databases (i.e., PsycTESTS and Psychological Testing Centre) and test publisher catalogs were surveyed. Examples of tests using each of the response formats are described below, with particular attention paid to the scoring procedures used.

PICK response format. One of the most common formats is the two-item unidimensional block. This format can be considered a FC normative measure because the two items of a block have different locations along the same trait, and only one of them is selected (Hicks, 1970). The block score is 1 when the more positive item is selected, or 0, otherwise. Test score is obtained by summing the block scores. For example, the test score for the *Preference for Solitude Scale* (Burger, 1995; see Table 1), which contains 12 two-item blocks, range from 0 to 12.

A second type of PICK format is the multidimensional two-item block. A block score of 1 is assigned to the dimension (i.e., item) selected, and 0 to the dimension not selected. The dimension score is the sum of the block scores. For example, Part 3 of Schuerger's (2008) *16PF-APQ test* (see Table 1) is comprised of 15 two-item blocks. Each item loads on one of six work activity preferences. Tests with more than two items per block have also been found. For example, the *Dellas Identity Status Inventory–Occupation* (Dellas & Jernigan, 1981) assesses the examinee's status in the occupational domain. The test contains 7 blocks of five items, each representing a possible status.

Some tests do not disclose relevant information, such as the scoring procedure or the polarities of the items forming a block (whether they are all positive, all negative, or mixed), but most do. The survey has found some innovative scoring and block composition schemes. Christiansen et al. (2005) proposed a scoring procedure for a test that measures conscientiousness and extraversion, comprised of 20 two-item blocks. Both items are either positive or negative, and only one of the items measures either conscientiousness or extraversion, whereas the other item measures an unrelated dimension. In a positive block, the dimension score is increased by 1 if the trait-related item is selected; in a negative block, 1 is added to the dimension score if the item unrelated to the trait, is selected. A block of this test is shown in Table 1.

MOLE response format. For the MOLE format, a block is typically comprised of three or four items. However, the number of items can be higher, as in *The Quest Profiler* (Pearce & Sik, 2000) and the *PAL-TOPAS* (Matthews & Oddy, 1997), which contain five- and six-item blocks, respectively. Blocks are scored by assigning 2 points to the (dimension tapped by the) most preferred item, 1 point to the unselected item/s, and 0 point to the least preferred item. Alternative and mathematically equivalent scoring schemes, such as (1, 0, –1), have also been proposed. Dimension scores are obtained as the sum of the block scores assigned to the dimensions. The first scoring scheme (2, 1, 0) is used in *The Occupational Personality Questionnaire (ipsative)* (OPQ32i; Brown & Bartram, 2009), and the *International Personality Item Pool–Multidimensional Forced Choice* (IPIP-MFC; Heggstad et al., 2006), both which have four-item blocks, and in the *Kuder Preference Record, Vocational* (Kuder-C; Kuder, 1988), the *Survey of Interpersonal Values* (Gordon, 2007), the *Survey of Personal Values* (Gordon, 2003), and the *Occupational Personality Questionnaire (ipsative shorted format)* (OPQ32r; Brown & Bartram, 2009), all of which have three-item blocks. The second scoring scheme, (1, 0, –1), is used in the *Gordon Personal Profile* and the *Gordon Personal Inventory* (GPP and GPI; Gordon, 2001). As an example, a block of the OPQ32i test is included in Table 1. However, the scoring procedures of some tests do not always fit this scheme. In the *DISCU Sipsative test* (Martinussen, Richardsen, & Varum, 2001), in an attempt to make the test less ipsative, not all the items selected are scored on every block. In the *PAL-TOPAS* personality questionnaire (Matthews & Oddy, 1997), the respondent must choose the two most and the two least applicable items, and the dimensions are scored by subtracting the number of *least* responses from the number of *most* responses.

RANK response format. Fewer tests were found using this format. Some examples are the *Canfield Learning Styles Inventory* (CLSI; Canfield, 1980), the *Concept Four* of the *Occupational Personality Questionnaires* (OPQ; Saville & Holdsworth Ltd, 1984), the *Study of Values* (SOV; Kopelman, Rovenpor, & Guan, 2003), and the *Learning Style Inventory* (Kolb & Kolb, 2005). These tests use four-item blocks and the respondents are asked to rank-order the four items to show their levels of agreement with the items. Blocks are scored by assigning the points 4, 3, 2, and 1 to the dimensions (items) ranked 1, 2, 3, and 4, respectively. The dimension score is obtained by summing the block scores of the dimension of interest. A block of the SOV test is shown in Table 1.

Item Response Theory (IRT) Models for FC Tests

To circumvent problems associated with ipsative measures, and extract scores from FC tests appropriate for inter-individual comparisons, a few models have been proposed within the IRT framework: the multi-unidimensional pairwise-preference (MUPP) model (Stark, Chernyshenko, & Drasgow, 2005), the McCloy et al. (2005) model, and the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011). In this study, the focus is turned to the MUPP model.

Stark et al. (2005) proposed an IRT model for a two-item block, from which respondents select the most preferred item. Typically, the two items measure different dimensions. If the model fits the data, and a suitable estimation procedure can be found, the estimated scores can be used to make inter-individual inferences (Chernyshenko et al., 2009). The response of a person to a block comprised of the items i and j will be denoted by Y_{ij} , where $Y_{ij} = 1$ if the respondent selects item i , and $Y_{ij} = 0$ if he or she selects item j . Let $\Pr(X_i = 1)$ be the probability of accepting the item i as a correct description. Stark et al. (2005) proposed that

$$\Pr(Y_{ij} = 1) = \frac{\Pr(X_i = 1) \Pr(X_j = 0)}{\Pr(X_i = 1) \Pr(X_j = 0) + \Pr(X_i = 0) \Pr(X_j = 1)}. \quad (1)$$

Stark et al. (2005) applied the *generalized graded unfolding model* (GGUM; Roberts, Donoghue, & Laughlin, 2000) for computing the item probabilities in Equation 1.

MUPP Extensions for the PICK, MOLE, and RANK Response Formats

The MUPP model can be extended to deal with the PICK, MOLE, and RANK response formats when there are more than two items in a block.

Model for the PICK format. Let $\Pr(A[A, B])$ be the probability of picking the item A when the two-item block $[A, B]$ is presented. $\Pr(A[A, B])$ is an alternative notation for $\Pr(Y_{ij} = 1)$ in Equation 1, and is a more convenient notation for blocks with more than two items. As in the two-item case, in a four-item block, $\Pr(A[A, B, C, D])$ is the probability of selecting item A out of the four possible items, and is given by

$$\Pr(A[A, B, C, D]) = \frac{\Pr(1, 0, 0, 0)}{\Pr(1, 0, 0, 0) + \Pr(0, 1, 0, 0) + \Pr(0, 0, 1, 0) + \Pr(0, 0, 0, 1)}, \quad (2)$$

where $\Pr(1, 0, 0, 0)$ is the joint probability of endorsing the item A but not the other three; $\Pr(0, 1, 0, 0)$ is the probability of endorsing the item B but not the other three, and so on. As in the MUPP model, (a) local independence is assumed to obtain the joint probabilities and (b) the GGUM is considered a proper model for the probability of endorsing each item. This is the PICK model, which can be applied when respondents have to choose exactly one of K items in a block.

Model for the RANK format. Luce (1959/2005) proposed a procedure to obtain the probability of a ranking of K elements based on the probabilities of picking the best from the whole set of K elements, then from the remaining $K - 1$, then from the remaining $K - 2, \dots$, and finally from the remaining last two elements. For example, the probability of the ranking (ACDB) is given by

$$\Pr(ACDB) = \Pr(A[A, B, C, D]) \times \Pr(C[B, C, D]) \times \Pr(D[B, D]). \quad (3)$$

In Equation 3, which can be referred to as the RANK model, $\Pr(A[A, B, C, D])$, $\Pr(C[B, C, D])$, and $\Pr(D[B, D])$ can be obtained by using the PICK model (i.e., Equation 2).

Model for the MOLE format. In the MOLE format, the probability of each block response can be obtained from the probabilities of the rankings of the items involved in the block. For example, in a block with four items A, B, C, and D, the probability of D and C being the most and least preferred items, $\Pr(D^{**}C)$, can be computed as the sum of the probabilities of all the possible rankings, where item D is ranked as the most preferred item and item C as the least preferred. In this example, the required probability would be

$$\Pr(D^{**}C) = \Pr(DABC) + \Pr(DBAC), \quad (4)$$

where $\Pr(DABC)$ and $\Pr(DBAC)$ can be obtained from Equation 3. This model will be referred to as the MOLE model.

Aim of the Article

The purpose of this article is to provide evidence on the type of information that test users may obtain when they use traditional scoring procedures for FC formats. A simulation study is designed to compare traditional scores, IRT estimates, and true scores, and know under which conditions, if any, the traditional scores could provide some useful information for inter-individual comparisons.

Method

Each simulated test measured four dimensions, and contained either two- or four-item blocks. In a four-item block, three response formats (PICK, MOLE, and RANK) were used; in a two-item block, only one item was picked (this format will be called PAIR). Items in a block tapped different dimensions so that each four-item block measured the same four dimensions. Similarly, each two-item block also measured two of the four dimensions. Two types of test scores were obtained for each respondent, one based on traditional scores and another on IRT estimates.

Scoring Procedures

The traditional scoring procedures described earlier were applied, but they were adapted to score blocks with both positive and negative items. For the PICK format, the selected dimension received 1 point when the item was positive and -1 point when it was negative; the non-selected dimensions received 0 point. This scheme would be equivalent to assigning 2, 0, and 1 points, respectively, if non-negative scores are preferred. For the MOLE format, the dimension of the *most* preferred item received 1 point when it is positive (-1 , if negative), and that of the *least* preferred -1 point when it is positive (1, if negative); the non-selected items provided 0 point to their dimensions. For the RANK format, if the item is positive, the most preferred dimension receives 4 points; the second most preferred dimension, 3 points; . . . and the least preferred, 1 point; if the item is negative, the most preferred receives 1 point; the second most preferred, 2 points; . . . and the least preferred, 4 points. In all formats, the test score on each dimension is the sum across blocks of the corresponding dimension scores.

IRT scores were obtained using the expected a posteriori (EAP) estimation procedure (Bock & Mislevy, 1982). It was deemed an appropriate procedure for estimating θ when, as in Stark et al. (2005) study, block parameters can be assumed to be known.

Simulation Study

Design. The following simulation factors were held constant: the number of dimensions the test measures was fixed at four; the number of respondents equaled 5,000, to improve results stability; and the tau parameters of the GGUM unfolding model were set to -1 . The following factors were varied: (a) block format: PAIR, PICK, MOLE, and RANK; (b) test length: 18 and 36 blocks, with the 18-block test used twice; (c) block discrimination (α parameter): low and high; (d) item polarity (sign of the δ parameters): all positive, half-and-half, or mixed blocks; and (e) variability of the δ parameters (standard deviation of the δ parameters of the items forming a block): low (0.2) and high (0.8); and (f) true θ correlation ρ : 0 and 0.5.

For the PICK, MOLE, and RANK formats, the four dimensions were measured in all the blocks; for the PAIR format, each dimension was measured in 9 or 18 blocks, and 3 or 6 blocks measured each of the six possible pairs of dimensions. The item α parameters were sampled from a $U(0.75-1.25)$ for low discrimination and $U(1.75-2.25)$ for high discrimination items. For the item polarity, in the all-positive condition, the δ parameters of all items ranged from 0.2 to 2; for the half-and-half condition, one half of the blocks had positive δ for all items and the other half had negative δ that ranged from -1 to 1; 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, with δ ranging from -1 to 1. The positive and negative δ 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. All δ parameters are sampled from uniform distributions. The δ parameters were generated repeatedly until they satisfied the variability requirement. Finally, the θ parameters were generated from a multivariate normal distribution with a mean vector of zero, variances of one, and a Pearson correlation between any two dimensions of ρ .

A total of 192 conditions are obtained by crossing the levels of the six factors: block format (4), test length (2), block discrimination (2), item polarity (3), variability of block localization (2), and true θ correlation (2). Each data set was scored using three procedures: (a) traditional scoring, (b) EAP estimation assuming $\rho = 0$, and (c) EAP estimation assuming $\rho = 0.5$. EAP estimates were obtained under the two different correlation assumptions to determine whether the congruence, or lack thereof, between the generation and estimation processes can affect the quality of the estimates.

Block parameters were treated as known when the EAP estimation procedure was applied. Ten quadrature nodes per dimension, equally spaced in a $(-3, 3)$ range, are used to obtain the EAP estimates of the latent trait vector θ . A preliminary study showed that using more quadrature points (i.e., 20) and a wider range of thetas from -4 to 4 produced no appreciable change in the EAP estimates for any of the models. Data generation and analysis were conducted by using the R v2.15.1 software (R Core Team, 2012).

Data generation process. For the PICK, RANK, and MOLE conditions, the probability of each of the 24 possible rankings of the four items was computed for each respondent and block, using Equation 3. Subsequently, one of the rankings was selected at random based on these probabilities and a multinomial distribution. The selected ranking served as the response for the RANK model. The response under the PICK and MOLE models were obtained from the item with the highest rank and the items with the first and fourth ranks, respectively. For the PAIR format, the MUPP probability in Equation 1 was calculated and compared with a uniform random number (0, 1) to simulate which of the two items of the block will be selected.

Measures of goodness of parameter recovery. Three measures (bias, root mean square error [RMSE], and the Pearson correlation between true and test scores) were computed to evaluate

the quality of the latent trait recovery. All the measures were computed for the IRT estimates, but only the Pearson correlation was computed to compare true thetas and traditional scores.

Results

Based on the overall results for traditional scores (first row of Table 2), the RANK model was the best performing model ($r = .67$), but it was only slightly better than the MOLE model ($r = .66$). The goodness of recovery of theta for these two models was very similar across all conditions. Table 2 shows that the difference across rows between the RANK and MOLE correlations was always less than .013. Therefore, under the conditions considered in the study, these models can be considered equivalent. Both were better than the PICK ($r = .58$) and PAIR ($r = .57$) models, which also showed very similar results (i.e., the absolute value of the difference across rows between the corresponding correlations is below .041).

For the EAP estimates, the RANK ($r = .93$; RMSE = .36) and MOLE ($r = .92$; RMSE = .36) models provide the best recovery of true thetas. Like in the traditional score case, the recovery efficiency of both formats was very similar across all conditions. These two models outperformed the PAIR ($r = .77$; RMSE = .62) and PICK ($r = .83$; RMSE = .53) models. However, unlike the pattern found for traditional scores, the PICK model was clearly better than the PAIR model when EAP estimation was involved. True thetas always correlated higher with EAP estimates than with traditional scores. The percentages of improvement in the correlations observed between EAP estimates and traditional scores were 34.5% (PAIR), 44.5% (PICK), 39.6% (MOLE), and 38.6% (RANK).

Concerning bias, the mean of EAP estimates practically equaled the mean of the true thetas. The differences of means found were very small: the overall difference is .0004 and those corresponding to the four models were .0002 (PAIR), $-.0005$ (PICK), .0009 (MOLE), and .0008 (RANK). Moreover, they do not seem to be systematically related to any of the simulation conditions. However, a conditional bias analysis revealed a slight and opposite-sign bias at extreme negative and positive true thetas, as the right graph of Figure 1 illustrates.

Concerning test length and block discrimination (see blocks and α s in Table 2), results showed that, as expected, the correlations between true and traditional scores were higher in the longer tests and more discriminating blocks conditions. This result also held for the EAP estimates. In both scoring schemes, increasing either the α parameters or the number of blocks had a similar impact on the recovery of theta. The mean correlations between true theta and the empirical scores were .769 and .712 (36 and 18 blocks), and .773 and .709 (higher and lower discrimination conditions). The mean increase in correlation, across models, produced by the 36-block test compared with the shorter test was .05 (7.5%) with traditional scoring, and .07 (8.4%) with EAP estimates; the increase due to the high block discrimination condition compared with the low block discrimination condition was .05 (7.7%) with traditional scoring, and .08 (10.1%) with EAP estimates. The impact of the increase of either test length or block discrimination is higher for the PAIR and PICK than for the MOLE and RANK models (see Blocks $\times$ α s in Table 2). For traditional scores, the combined effect of high block discrimination and 36-block test yielded a moderately good recovery: $r = .63$ (PAIR), $r = .62$ (PICK), $r = .69$ (MOLE) and $r = .69$ (RANK); for the same condition, parameter recovery was very good when EAP estimation was applied: $r = .96$ and RMSE = .27, $r = .96$ and RMSE = .27, $r = .92$ and RMSE = .38, and $r = .87$ and RMSE = .48, for the RANK, MOLE, PICK, and PAIR models, respectively.

These findings are consistent with well-known IRT results: higher discrimination and a larger number of blocks increase test information and should lead to better θ estimates. The differences observed in the four models can be attributed to the fact that they differ in the amount of

Table 2. Correlations Between True Thetas and Traditional Scores or EAP Estimates.

	Traditional scores				EAP			
	PAIR	PICK	MOLE	RANK	PAIR	PICK	MOLE	RANK
Overall	.571	.575	.662	.669	.768	.831	.924	.927
Blocks								
18	.539	.547	.646	.654	.716	.786	.903	.906
36	.603	.602	.678	.683	.820	.876	.945	.947
α s								
(0.75, 1.25)	.535	.549	.647	.654	.708	.777	.897	.902
(1.75, 2.25)	.607	.600	.677	.684	.828	.885	.950	.951
Blocks $\times$ α s								
18								
(0.75, 1.25)	.496	.513	.624	.633	.648	.720	.865	.860
(1.75, 2.25)	.582	.581	.668	.675	.785	.852	.940	.941
36								
(0.75, 1.25)	.575	.585	.669	.674	.768	.834	.931	.931
(1.75, 2.25)	.631	.619	.687	.692	.872	.919	.961	.962
δ s								
Polarity								
All positive	.489	.506	.575	.581	.697	.738	.884	.887
Half-and-half	.529	.553	.636	.639	.769	.838	.934	.937
Mixed	.695	.665	.775	.786	.838	.873	.953	.955
Variability								
.2	.586	.576	.652	.657	.746	.804	.908	.908
.8	.556	.573	.672	.680	.790	.858	.940	.942
Polarity $\times$ Variability								
All positive								
.2	.505	.497	.555	.559	.662	.733	.851	.856
.8	.473	.514	.596	.603	.733	.833	.917	.919
Half-and-half								
.2	.541	.556	.630	.632	.738	.809	.920	.924
.8	.517	.550	.641	.646	.800	.866	.949	.950
Mixed								
.2	.710	.677	.771	.780	.839	.870	.952	.954
.8	.679	.654	.779	.791	.837	.876	.955	.956
θ s								
Generated								
0	.666	.665	.767	.771	.765	.823	.928	.930
.5	.476	.484	.557	.566	.723	.802	.906	.909
Estimated								
0					.743	.812	.916	.919
.5					.745	.812	.918	.920
Generated vs. estimated								
0								
0					.788	.841	.934	.936
.5					.742	.804	.921	.923
.5								
0					.697	.783	.898	.902
.5					.748	.821	.914	.917

Note. EAP = expected a posteriori.

involvement expected of the respondents. The simulation study showed that more elaborate response formats yield better θ recovery indices.

Figure 1. True scores and traditional scores (left) or EAP estimates (right) scatterplots.

Note. EAP = expected a posteriori.

Results also showed that tests including items with larger differences among their δ s (either in *polarity* or *variability*) do better than tests with more similar δ s. For both traditional scoring and EAP estimates, the mixed condition outperformed the other two conditions, and the half-and-half condition outperformed the all-positive condition. In traditional scoring, it is noteworthy that for all response formats correlations reached their highest values (about .70, see Polarity of δ s in Table 2) in the more heterogeneous polarity (mixed) condition. However, it can also be seen that more variable δ s ($SD = .8$) improved the recovery when EAP estimation was used, but it had no clear effect on traditional scoring. Moreover, it appeared that the δ polarity produced a more substantial effect than the δ variability for the levels of both factors examined in this article.

For the trait correlation (lower portion of Table 2), θ recovery was better, for both traditional and IRT scores, when θ s were generated independently than when they were generated dependently. The improvement was particularly notable for the traditional scores. The observed increases in the correlation (and corresponding percentages) were .19 (39.9%), .18 (37.3%), .21 (37.8%), and .21 (36.3%) for the PAIR, PICK, MOLE, and RANK models, respectively. For the EAP estimates, the corresponding numbers were .04 (5.5%), .02 (2.5%), .02 (2.2%), and .02 (2.1%). In the independent traits case, θ recovery correlations can be improved and reach acceptable to high levels of recovery when combined with the large number of blocks, high discrimination parameters, and high heterogeneous polarity conditions: $r = .80$ (PAIR), $r = .77$ (PICK), $r = .87$ (MOLE), and $r = .88$ (RANK), for the traditional scores; and $r = .92$ (PAIR), $r = .95$ (PICK), $r = .98$ (MOLE), and $r = .98$ (RANK), for the EAP estimates.

It can also be observed that the θ recovery was better, as indicated by both the correlation and RMSE indices, when the correlation value used to obtain the EAP estimates matched that correlation used for data generation. Across models, the mean correlation EAP and true thetas were .86 and .83 when there was a match and there was no match, respectively (see generated vs. estimated θ s in Table 2). When the estimation assumed that the traits were correlated, the recovery efficiency was basically the same when the data were generated as independent or correlated (across models, the mean correlation between EAP and true thetas was .85 for both conditions). In contrast, assuming in the estimation process that the traits were independent had greater impact—the correlation was .88 if there was match, and only .82 if there was not. These trends were also observed with the RMSE.

The mean correlations between the true scores with traditional scores and EAP estimates were .62 and .86, respectively. The .24 difference was expected to be related to the specific characteristics of the model used in the data generation. Figure 1 shows the scatterplots of the

true and the traditional scores (left graph) and the EAP estimates (right), for one dimension of one of the simulated conditions. Pearson correlations were .80 and .92, respectively. As indicated, the authors used the GGUM as the model to obtain the probabilities of endorsing each block item. As an unfolding model, the probability of endorsing the item decreases as the distance between the true theta and the item location parameter increases. The curvilinear trend of the left graph of Figure 1 could be related to the unfolding model used in the data generation. For example, very high true thetas would be very likely far away from the item location and yield a probability of endorsing the item that is lower than the probability associated with a different process (e.g., dominance model). However, based on the good properties of EAP estimators (Bock & Mislevy, 1982), the EAP estimates were expected to show (and previous results of the present study confirmed) a small bias across most of the range of true thetas and, therefore, a linear relationship with true thetas. The right graph of Figure 1 shows that this prediction was observed. For this reason, the authors conjecture that the data generation process is partly responsible for the difference found in the correlations between the true theta and the traditional scores and EAP estimates.

The Spearman correlation, an indicator of the agreement between two rankings and less affected by non-linearity than Pearson correlations, has been applied as well to the data set to gain additional information on the role of non-linearity in the results shown in Table 2. The means across response formats of Spearman correlations between the true scores with traditional and EAP estimates were 0.66 and 0.85, respectively. So, the .24 gap found earlier now reduces to .19, due mainly to the increase witnessed in the traditional scores correlation. This result strengthens the idea that a non-linear relationship may be to some extent behind the underperformance of traditional scores. In this same vein, for traditional scores, but not so for EAP estimates, the difference between the PICK and PAIR Pearson correlations was negligible (see first row of Table 2): The differences between these two formats were 0.004 and 0.063 for traditional and EAP scores, respectively. When computing Spearman correlations both differences increase to 0.056 and 0.073, being now more similar, and showing, as expected, higher correlations for the PICK condition under both scoring schemes.

Finally, ipsative or partially ipsative traditional scores were expected to be negatively correlated despite a true correlation of zero between the true thetas. For the particular condition shown in Figure 1, the mean and range of the correlations between the traditional scores observed across the four dimensions are $-.258$ and $(-.288, -.214)$; for the EAP estimates, these values were $-.023$ and $(-.035, -.015)$, which were closer to the observed correlations between the generated thetas: $-.006$ and $(-.025, .008)$.

Discussion

This article explored some of the most widely used traditional scoring procedures for FC formats, and examined the conditions under which these procedures could rank-order the respondents as IRT estimates would. To the authors' knowledge, studies that evaluate systematically the quality of the most common traditional scoring schemes for FC test formats by comparing empirical scores with their true scores have not been carried out. This work filled this gap by making such a comparison possible through a simulation study. Three extensions of the MUPP model (PICK, MOLE and RANK) were proposed to handle widely used FC formats. By using these extensions, true scores, EAP estimates, and traditional scores can be simultaneously obtained, making their comparison possible.

In general, EAP estimates perform appreciably better than traditional scores. The obtained results indicated a strong improvement for IRT modeling under the explored conditions, which were related to block format, test length, block discrimination, delta polarities, intrablock delta

variability, and trait correlations. The obtained θ recovery indices for the EAP estimates could also be considered as an upper limit that allows us to show the efficiency of recovery lost when traditional scoring is used instead of EAP estimates for each particular condition.

The current study suggests that the RANK format works better than other formats both for IRT estimates and traditional scores. This finding is not unexpected, because the RANK response format capitalizes on more and better information. However, the RANK format is not substantially better than the MOLE format. This finding may depend on the number of items each block contains. Obviously, in a test with three items per block, selecting the best and worse items and rank-ordering all the items are equivalent responses. The use of just four items involves the minimal difference between the two response formats, and the results seem to suggest that the additional information gained by ranking the two non-selected options under the MOLE format does not represent any practical advantage. More than four items should therefore be explored to determine the potential benefits of the RANK over the MOLE format. For the four-item blocks, the current practical recommendation is to use the MOLE rather than the RANK format because the MOLE task is easier for the respondents, and no substantial loss in recovery can be expected.

Concerning the number of items to include in each block, for the EAP estimates, the results suggest that it is better to use four (PICK format) than two (PAIR format) items. The advantage is due to the additional information provided by the former compared with the latter. In this study, blocks in the PICK format condition included items from each of the four dimensions, whereas in the PAIR case, from two dimensions only. Given that in some of the explored conditions differences are small, in practical situations, the balance between the complexity of the task (e.g., length, time, cost) and the expected improvement in accuracy of a model over another should be considered. However, for traditional scores, it is preferable to use the PAIR format, because the results would be very similar. However, as it has been shown, the correlation of traditional scores with true thetas is higher for the PICK condition than the PAIR if Spearman instead of Pearson correlations are computed, indicating that non-linearity may be an issue in this comparison.

The current study also showed that a FC test should have a large number of blocks and highly discriminating items covering well the levels of the traits considered to obtain high levels of theta recovery when using IRT estimates (this was not as obvious with traditional scoring). The statement adheres to the general principle that developing good test instruments is a prerequisite to obtaining high-quality scores. It should also be underscored that a FC test works better with balanced blocks (i.e., mixed blocks), and independent rather than correlated dimensions. Traditional scoring provided its best results under these conditions; therefore, these are also the conditions under which the use of traditional scoring could be recommended. In these conditions, acceptable correlations can be expected.

Additional work needs to be carried out to better understand the properties of the proposed methods. First, although the superiority of the RANK and MOLE models over the PICK and PAIR models has been found, the results are entirely based on simulated data. It is not clear whether the relative efficiency would be maintained with real data. The RANK and MOLE formats advantage is primarily due to the fact that they extract more information from the respondent on each block compared with the PICK format. Although these two formats are more informative, they can also be more taxing. So, the substantial advantage of the RANK and MOLE over the PICK model may diminish or even disappear when real respondents are involved.

The advantages of the EAP estimates over their traditional score counterparts should not come as a complete surprise because obtaining the EAP estimates require more information. Specifically, they required item parameter measured without error as input. Future studies

should investigate whether the EAP estimates will retain their advantage when item parameters need to also be estimated. De la Torre, Ponsoda, Leenen, and Hontangas (2011) obtained preliminary results showing that estimating some item parameters instead of treating them as known still provided better recovery results than those obtained with traditional scores.

Some limitations of this study should be acknowledged. First, an unfolding model (i.e., GGUM) is used for data generation and, as noted above, the results may be affected by this particular type of model. Future research should check whether the conclusions change if the GGUM is replaced by a dominance model. Second, in this study, each test measures just four dimensions, although many personality tests with a large number of scales, like 16PF or OPQ, are available. When using traditional scores, the dependence among dimension scores diminishes as the number of dimensions increases (Baron, 1996; Clemans, 1966). Therefore, a higher number of dimensions should be tried to check whether the witnessed IRT versus traditional scoring differences progressively vanish as test dimensionality increases. Third, alternative scoring procedures based on compositional data analysis have been proposed to analyze ipsative data (Aitchison, 2003; de Vries & van der Ark, 2008; van Eijnatten et al., 2015). A comparison of IRT-based scores against alternative scoring strategies used in compositional data analysis deserves exploration as well.

Finally, traditionally scored FC tests are frequently used to extract normative information. Therefore, this study intended to examine under which conditions the normative information they may extract is more likely to be correct. However, although correlations between traditional scores and true scores can be moderate, traditional scoring methods are not advised because they face well-known problems (related to reliability, factorial structure, and predictive validity) that can be solved when adequate IRT models are applied.

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: The research has been funded by the Spanish Ministry of Economy and Competitiveness, Project PSI2012-33343.

References

- Aitchison, J. (2003). *The statistical analysis of compositional data* (Reprint). Caldwell, NJ: Blackburn.
- 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.
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, 14, 317-335.
- 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). *Development and psychometric properties of OPQ32r* (Supplement to the OPQ32 technical manual). Thames Ditton, UK: SHL Group Limited.
- 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.
- Burger, J. M. (1995). Individual differences in preference for solitude. *Journal of Research in Personality*, 29, 85-108.
- Canfield, A. A. (1980). *Learning styles inventory: Manual*. Ann Arbor, MI: Humanics Media.
- Chernyshenko, O. S., Stark, S., Prewett, M. S., Fray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance*, 22, 105-127.
- 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.
- Clemons, 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.
- Dellas, M., & Jernigan, L. P. (1981). Development of an objective instrument to measure identity status in terms of occupation crisis and commitment. *Educational and Psychological Measurement*, 41, 1039-1050.
- de Vries, A. L. M. (2006). *The merit of ipsative measurement: Second thoughts and minute doubts* (Doctoral thesis). Maastricht University, Maastricht, The Netherlands.
- de Vries, A. L. M., & van der Ark, L. A. (2008, March 31). *Scoring methods for ordinal multidimensional forced-choice items*. CODAWORK'08, Girona, Spain.
- Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York, NY: Dryden Press.
- Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories*. New York, NY: Holt, Rinehart & Winston.
- Gordon, L. V. (2001). *PPG-IPG. Perfil e inventario de personalidad* [Gordon personal profile and Gordon personal inventory]. Madrid, Spain: TEA Ediciones.
- Gordon, L. V. (2003). *SPV. Cuestionario de valores personales* [Survey of personal values]. Madrid, Spain: TEA Ediciones.
- Gordon, L. V. (2007). *SIV. Cuestionario de valores interpersonales* [Survey of interpersonal values]. Madrid, Spain: TEA Ediciones.
- 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.
- 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 spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61, 153-162.
- Kolb, A., & Kolb, D. (2005). *The Kolb Learning Style Inventory—Version 3.1* (Technical specifications). Boston, MA: Hay Resource Direct.
- Kopelman, R. E., Rovenpor, J. L., & Guan, M. (2003). The study of values: Construction of the fourth edition. *Journal of Vocational Behavior*, 62, 203-220.
- Kuder, G. F. (1988). *Kuder preference record: Vocational (Kuder-C)* (Manual). Madrid, Spain: TEA Ediciones.

- Luce, R. D. (2005). *Individual choice behavior*. Mineola, NY: Dover Publication. (Original work published 1959).
- Martinussen, M., Richardsen, A. M., & Varum, H. W. (2001). Validation of an ipsative personality measure (DISCUS). *Scandinavian Journal of Psychology*, 42, 411-416.
- 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., Heggstad, 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.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.
- Pearce, A., & Sik, G. (2000). *The quest profiler*. Norfolk, VA: Eras, expertise at work. Retrieved from <http://www.eras.co.uk/psychometrics/the-quest-profiler-tm/>
- R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org/>
- 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 & Holdsworth Ltd. (1984). *Occupational Personality Questionnaires manual*. Esher, UK: Author.
- 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.
- Schuerger, J. M. (2008). *16PF-APQ (16 PF-Adolescent Personality Questionnaire) manual*. Madrid, Spain: TEA Ediciones.
- 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.
- Travers, R. M. W. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin*, 48, 62-70.
- 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 & Quantity*, 49, 559-579. doi:10.1007/s11135-014-0009-8