Journal of Work and Organizational Psychology (2019) 35(2) 75-83

Journal of Work and Organizational Psychology

https://journals.copmadrid.org/jwop

The Journey from Likert to Forced-Choice Questionnaires: Evidence of the Invariance of Item Parameters

Daniel Morillo[a], Francisco J. Abad[b], Rodrigo S. Kreitchmann[b], Iwin Leenen[c], Pedro Hontangas[d], and Vicente Ponsoda[b]

[a]*Instituto de Ingeniería del Conocimiento, Madrid, Spain;* [b]*Universidad Autónoma de Madrid, Spain;* [c]*Universidad Nacional Autónoma de México, Mexico;* [d]*Universidad de Valencia, Spain*

ARTICLE INFO

Article history:
Received 15 January 2019
Accepted 29 March 2019
Available online 21 June 2019

Keywords:
Forced-choice questionnaires
Invariance
IRT
Likelihood Ratio test
MUPP-2PL

A B S T R A C T

Multidimensional forced-choice questionnaires are widely regarded in the personnel selection literature for their ability to control response biases. Recently developed IRT models usually rely on the assumption that item parameters remain invariant when they are paired in forced-choice blocks, without giving it much consideration. This study aims to test this assumption empirically on the MUPP-2PL model, comparing the parameter estimates of the forced-choice format to their graded-scale equivalent on a Big Five personality instrument. The assumption was found to hold reasonably well, especially for the discrimination parameters. In the cases in which it was violated, we briefly discuss the likely factors that may lead to non-invariance. We conclude discussing the practical implications of the results and providing a few guidelines for the design of forced-choice questionnaires based on the invariance assumption.

El viaje desde los cuestionarios Likert a los cuestionarios de elección forzosa: evidencia de la invarianza de los parámetros de los ítems

Palabras clave:
Cuestionarios de elección forzosa
Invarianza
IRT
Test de razón de verosimilitudes
MUPP-2PL

R E S U M E N

Los cuestionarios de elección forzosa multidimensionales son bastante apreciados en la literatura de selección de personal por su capacidad para controlar los sesgos de respuesta. Los modelos de TRI desarrollados recientemente normalmente asumen que los parámetros de los ítems permanecen invariantes cuando se emparejan en bloques de elección forzosa, sin dedicarle mucha atención. Este estudio tiene como objetivo poner a prueba empíricamente este supuesto en el modelo MUPP-2PL, comparando las estimaciones de los parámetros del formato de elección forzosa con su equivalente en escala graduada, en un instrumento de personalidad *Big Five*. Se encontró que el supuesto se cumplía razonablemente bien, especialmente para los parámetros de discriminación. En los casos en los que no se cumplió se discuten brevemente los posibles factores que pueden dar lugar a no invarianza. Concluimos discutiendo las implicaciones prácticas de los resultados y proponiendo algunas pautas para el diseño de cuestionarios de elección forzosa basados en el supuesto de invarianza.

Several meta-analyses have shown that the Five-Factor Model of personality predicts a wide range of performance outcomes in the workplace and is a useful framework for organizing most personality measures (see, e.g., Barrick & Mount, 1991; Judge, Rodell, Klinger, Simon, & Crawford, 2013; Tett, Rothstein, & Jackson, 1991). Conscientiousness and emotional stability are consistently found to predict job performance for all occupations. The other three dimensions are valid predictors for specific criteria and occupations (Barrick, Mount, & Judge, 2001; Salgado, Anderson, & Tauriz, 2015).

Most personality questionnaires use single-stimulus items (e.g., Likert type). Forced-choice questionnaires (FCQs) are another type of psychological measurement instruments used in the evaluation of non-cognitive traits, such as personality, preferences, and attitudes (see, e.g., Bartram, 1996; Christiansen, Burns, & Montgomery, 2005; Ryan & Ployhart, 2014; Saville & Willson, 1991). From recruitment and selection professionals' point of view, the main interest for these instruments is their ability to control for certain responses biases. Evidence suggests they are comparatively robust against impression management attempts, which may easily arise in high-stakes contexts such as a selection process. Impression management has at least three effects on personality questionnaire scores: (1) a decrease of their reliability index, (2) lower validity, and (3) an alteration of the

Cite this article as: Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Journal of Work and Organizational Psychology, 35*, 75-83. https://doi.org/10.5093/jwop2019a11 [Antonio García-Izquierdo and David Aguado were the guest editors for this article].

Funding: This research is funded by the Spanish government's Ministerio de Economia y Competitividad, projects PSI 2015-65557-P and PSI 2017-85022-P.
Correspondence: daniel.morillo@iic.uam.es (D. Morillo).

individual rankings. These effects are, of course, especially relevant in the domain of personnel selection, as they affect hiring decisions negatively (Salgado & Lado, 2018).

Despite their resistance to faking, FCQs would not be relevant if they did not fare well in performance prediction when compared to alternative assessment formats. In recent years, a few meta-analyses (Salgado, 2017; Salgado et al., 2015; Salgado & Tauriz, 2014) have examined the predictive validity of FCQs and compared it with single-stimulus questionnaires. The evidence is that FCQs producing quasi-ipsative scores (described later) fulfill better the abovementioned criterion validity requirement.

Multidimensional FCQs are a special case that prompts the examinee to choose among stimuli (i.e., items) that are valid against different criteria (Ghiselli, 1954). In contrast, unidimensional FCQs prompt to choose among items that are valid against the same criterion, or among valid and invalid (i.e., "suppressor") items (Sisson, 1948). Although the suppressor-item format was the original proposal, the multidimensional format rapidly imposed itself, due to its ability to assess several traits simultaneously (Scott, 1968).

Depending on the method used to score the traits, multidimensional FCQs may yield ipsative or quasi-ipsative scores (Hicks, 1970). A respondent's task can be organized to avoid a constant total sum of the measured dimensions. This fact may be achieved, for example, by requesting the candidates not just to select the item that describes them best or worse, but also to rate how good (or bad) the description is.

"Strong" ipsativity implies that the total sum of scores in a multidimensional FCQ is fixed (Hicks, 1970). Ipsative scores violate the assumptions of the Classical Test Theory, leading to a distortion in reliability and construct validity. From a practical point of view, this implies an impossibility to compare persons according to their level on the traits assessed (Cornwell & Dunlap, 1994). Ipsativity issues have led to a great controversy revolving around the forced-choice (FC) format. However, this controversy has largely ignored the fact that ipsativity is a property of the direct scoring method, not of the format itself. The confusion may probably stem from the assimilation of both terms (see, e.g., Christiansen et al., 2005; Saville & Willson, 1991, for examples of the use of the "ipsative" term as a synonym of "forced-choice").

Some researchers have proposed using Item Response Theory (IRT) models to circumvent the problems of ipsativity. These would allow obtaining normative scores on the trait levels. Although the literature does not provide examples so far, they may also help developing scoring methods, such as item weightings for computing quasi-ipsative scores.

The multi-unidimensional pairwise preference (MUPP; Stark, Chernyshenko, & Drasgow, 2005) was the first IRT model to be proposed for multidimensional FCQs; it is characterized mainly by (1) applying to two-item blocks and (2) by assuming that each item's measurement model is an "ideal point" model, which implies that the probability of agreeing with a response option decreases with the "distance" of the respondent to the item location on the trait continuum. The MUPP model also assumes that the response process implies independent decisions between the two options (Andrich, 1989, 1995). This assumption leads, in turn, to hypothesize that item parameters do not change when those items are paired in FC blocks. This assumption is paramount for the validity of multidimensional FC instruments and, as we will explain below, is the focus of this paper.

The Thurstonian IRT model (TIRT; Brown & Maydeu-Olivares, 2011), based on Thurstone's (1927) law of comparative judgment, followed in chronological order. Unlike the MUPP model, it applies to blocks with more than two items. It also assumes a "dominance" rather than ideal point measurement model—the probability of agreement with each item increases (or decreases) monotonically with a respondent's latent trait score.

The MUPP-2PL model (Morillo et al., 2016) is a variant of the MUPP; as such, it applies to two-item blocks as well. It differs from the original MUPP in that the probability of agreement with each response option, as in the Thurstonian IRT model, is modeled by a dominance function. More precisely, it assumes that a 2-parameter logistic (2PL; Birnbaum, 1968) curve models the response to each of the two items. This curve is expressed as

$$\mathrm{P}\left(X_{i_p j}=1\middle|\boldsymbol{\theta}_j\right)=\phi_L\left(a_{i_p}\left(\theta_{\tilde{i_p}j}-b_{i_p}\right)\right) \quad (1)$$

where $\phi_L(\cdot)$ is the logistic function, $\theta_{\tilde{i_p}j}$ the latent trait tapped by the item i_p, and a_{i_p} and b_{i_p} its characteristic parameters. These can be interpreted, respectively, as a "discrimination" parameter and a "location" parameter: a_{i_p} would indicate how sensitive or discriminant i_p is to differences in $\theta_{\tilde{i_p}j}$ b_{i_p} would be a "point of indifference" in $\theta_{\tilde{i_p}j}$ (i.e., where both the probability of agreeing and disagreeing with the item would be equal to .50).

The assumptions of the MUPP-2PL model imply that if item i_p was presented independently in a dichotomous response format, the probability that a respondent agreed with it would also be given by Equation 1 (see Morillo, 2018). Thus, if a FC block consists of items i_1 and i_2 and we presented those same items in a dichotomous format, their parameters and the FC block parameters should be equivalent. The response function of a bidimensional FC block (i.e., a block with two items tapping different latent dimensions) is given by

$$\mathrm{P}_i\left(Y_{ij}=1\middle|\boldsymbol{\theta}_j\right)=\phi_L\left(a_{i_1}\theta_{\tilde{i_1}j}-a_{i_2}\theta_{\tilde{i_2}j}+d_i\right) \quad (2)$$

with

$$d_i=a_{i_2}b_{i_2}-a_{i_1}b_{i_1} \quad (3)$$

Therefore, a_{i_1} and a_{i_2} should be the same for the dichotomous items and the FC block, while d_i should be a linear combination of the two location parameters b_{i_1} and b_{i_2} in the dichotomous format. We call this the "invariance assumption", as it implies that the parameters are invariant to both the format (FC versus dichotomous) and the within-block context (i.e., the other item(s) a certain item is paired with).

Previous research has not subjected this assumption to abundant scrutiny; on the contrary, giving it for granted is prevalent in the literature (see, e.g., Stark et al., 2005). Lin and Brown (2017) performed a retrospective study on massive data from the Occupational Personality Questionnaire (OPQ; Bartram, Brown, Fleck, Inceoglu, & Ward, 2006). Applying the Thurstonian IRT (Brown & Maydeu-Olivares, 2011) model, they compared the parameters in two versions of the instrument: OPQ32i, which uses a partial-ranking task with four items per block (most/least like me), and OPQ32r, a reviewed version that dropped one item from each block (Brown & Bartram, 2011), and implied a complete-ranking task with three items per block. They found that the parameters largely fulfilled the invariance assumption. When it was not fulfilled, they also identified possible causal factors. They interpreted them as within-block context effects—variations of the item parameters due to the other item(s) in the same block. However, they did not compare the FC format with the items individually applied in a single-stimulus format. It is also noteworthy that the OPQ Concept Model is not the most widely accepted theoretical model of personality.

Testing the invariance assumption is crucial for the design of FCQs. When designing such an instrument, a practitioner may be tempted to rely upon the item parameter estimates without further consideration. However, if the invariance assumption is violated, the FC block parameters may change dramatically, leading in turn to a lack of construct validity in the latent trait scores. The purpose of this study is thus to test the invariance assumption of the MUPP-2PL model. In order to this, we will compare the parameters of a set of

items designed to measure personality variables, applying them as bidimensional FC blocks, and "individually".

Single-stimulus items are usually applied in a graded-scale (GS) format. Therefore, we first propose a method for testing the invariance assumption with this presentation format. Then, we apply this method to an empirical dataset assessing the Big Five construct domain. Finally, we discuss the application of the method and our results, giving some guidelines about their consequences for the design of FCQs.

A Test of the Invariance Assumption with Graded-scale Items and Forced-choice Blocks

The traditional format of presenting non-cognitive items in questionnaires is the GS or Likert format. This format implies a series of responses graded in their level of agreement with the item statement. Compared to the dichotomous format, the additional categories in the response scale provide a surplus of information that yields more reliable latent trait scores (Lozano, García-Cueto, & Muñiz, 2008).

The Graded Response Model (GRM; Samejima, 1968) can be applied to the response data from a GS questionnaire. According to the GRM, if a person responds to an item i_p that has $m+1$ categories (from 0 to m), they will choose category k or higher (with k being a category from 1 to m) with probability

$$\mathrm{P}\left(X_{i_p j} \geq k \middle| \boldsymbol{\theta}_j\right) = \phi_L\left(a^*_{i_p}\left(\theta_{\widetilde{i_p} j} - b_{i_p k}\right)\right) \quad (4)$$

where $a^*_{i_p}$ is the discrimination parameter in GS format, and $b_{i_p k}$ is the location parameter for category k. This latter parameter represents the point in the latent trait continuum where the probability of agreeing with i_p at least as much as stated by k equals .5.

When $m = 1$ there are two response categories, and Equation 4 is reduced to the 2PL model expressed in Equation 1 with $b_{i_p 1} = b_{i_p}$. When $m > 1$, we may consider a recoding of the responses, for a given arbitrary category k' between 1 and m (both included), such that the new value is 1 if the response is equal to k' or higher, and 0 otherwise. This recoding implies a representation of the responses to the Likert-type items as a dichotomous format, with a response probability given by Equation 1. According to the GRM, when dichotomizing a GS item, its parameters are expected to remain unchanged (Samejima, 1968). Therefore, we can assume the parameters of a bidimensional FC block to be equivalent to the parameters of its constituent items, as expressed in Equation 2.

We must make a caveat here, since none of $b_{i_p k'}$ parameters can be considered equivalent to the actual b_{i_p}. As we stated before, the latter represents the point in the latent trait continuum where $\mathrm{P}\left(x_{i_p j} = 1 \middle| \boldsymbol{\theta}_j\right) = .5$ in Equation 1, when such a statement is presented as a dichotomous item. When we perform a dichotomization of a GS format as stated above, the k' threshold category chosen does not necessarily imply that its $b_{i_p k'}$ parameter coincides with the parameter from the dichotomous presentation as in Equation 1. We consider however that the equivalence given between the dichotomized GRM and the 2PL models justify considering and assessing the linear combination of the item category location parameters as a proxy for the block intercept parameter.

In conclusion, testing the invariance assumption of the MUPP-2PL in a bidimensional FC block implies testing three hypotheses of equality of parameters: of the discrimination parameters of the two items ($a_{i_1} = a^*_{i_1}$ and $a_{i_2} = a^*_{i_2}$), and of the block intercept parameter with the correspondent linear combination of the item parameters ($d_{ik'} = a^*_{i_1} b_{i_2 k'} - a^*_{i_2} b_{i_1 k'}$), which can be performed on the m values of k'. These can be done by means of a likelihood ratio test (Fisher, 1922), comparing an unconstrained model with the nested, constrained one, applying the corresponding restriction of equality. As it is well known, the resulting test statistic is asymptotically chi-square distributed under the null hypothesis (Wilks, 1938), in this case with one degree of freedom. This enables a very simple procedure for testing the invariance assumption, based on a well-known and reliable methodology. In order to put this method to test, and provide evidence regarding the invariance assumption, the following section exemplifies the application and results of this method.

Method

Materials

We used a dataset consisting of responses to a GS questionnaire and a multidimensional FCQ. Both instruments shared a large number of items and were answered by a common group of participants, so they were suitable to apply the invariance assumption tests. The contents of this dataset are described below.

Instruments

Graded-scale questionnaire. It consisted of 226 GS items presented in a five-point Likert scale (*completely disagree – disagree – neither agree nor disagree – agree – completely agree*); there were $m = 4$ category thresholds therefore. The items were designed to measure the dimensions of the Big Five model (McCrae & John, 1992). An example of an emotional stability item is as follows: "Using the previous five-point scale, indicate your agreement with this statement: 'Seldom feel blue'." Example statements for the other four dimensions are these: "Make friends easily" (extraversion), "Have a vivid imagination" (openness to experience), "Have a good word for everyone" (agreeableness) and "Am always prepared" (conscientiousness). The five items are selected from the Big-Five IPIP Inventory item pool (Goldberg, 1999).

Forty-four items were applied for each of the five traits. One hundred twenty-two of these items were direct (i.e., positively keyed), and 98 were inverse (i.e., negatively keyed; see Morillo, 2018); polarity was aimed to be balanced among the different traits, with 22 to 26 direct items and 18 to 22 inverse items per trait. The remaining six items were directed items (e.g., "Select the *disagree* response"), applied to control the quality of each participant's responses (Maniaci & Rogge, 2014). The items were distributed in two booklets, with 113 items each, with the directed items at positions 26, 57, and 88 and 23, 55, and 87 in the first and second booklet, respectively.

Table 1. Distribution of the FC Blocks by Trait

	Item 2				
Item 1	Emotional Stability	Extraversion	Agreeableness	Openness	Conscientiousness
Emotional Stability	-	3	3	3	3
Extraversion	5	-	5	3	5
Agreeableness	4	3	-	4	5
Openness	5	5	4	-	4
Conscientiousness	5	3	3	4	-

Table 2. Summary of Results of the Invariance Assumption Tests

Parameters	Estimate Statistics			Non-invariant Parameters	
	Correlations	Mean error	MRE	Count	%
Discrimination	.93	-0.13	-0.21	2	2.33
Intercept (threshold 1)	.87	0.78	-0.34	13	33.33
Intercept (threshold 2)	.91	0.71	-0.45	15	38.46
Intercept (threshold 3)	.94	0.52	-0.08	12	30.77
Intercept (threshold 4)	.91	0.11	-0.16	10	25.64

Note. MRE = mean relative error; RMSE = root mean square error.

Table 3. Likelihood Ratio Test Statistics of the Constrained Models

Block	Trait		Polarity		Discrimination		Intercept			
	1	2	1	2	Item 1	Item 2	Category 1	Category 2	Category 3	Category 4
2	ES	Ag	+	-	4.27	0.04	4.78	0.97	1.83	0.02
3	Co	ES	+	+	0.17	13.57	22.85[1]	20.25[1]	0.78	26.15[1]
4	ES	Ag	+	-	0.70	0.70	0.02	1.56	0.00	0.85
5	Ag	Op	+	-	5.69					
6	Op	ES	+	-	3.36	0.00	22.01[1]	43.79[1]	2.87	6.67
7	ES	Op	+	+	0.10	0.01	2.62	2.30	7.62	23.67[1]
10	ES	Co	+	+	1.04	0.65	13.64	7.95	8.71	15.03[1]
12	Op	Co	+	+	0.10	0.00	0.47	3.44	1.21	4.47
13	Co	Op	+	+	1.35	4.22	3.49	17.28[1]	14.78[1]	11.23
14	Ag	Co	+	+	8.30	6.53	0.88	2.01	0.30	0.02
15	Ag	Co	+	-	12.82					
16	Op	Ag	+	+	0.00	1.04	2.87	2.10	9.13	9.79
19	ES	Co	+	-	7.89					
20	Op	ES	+	-	0.17	1.57	7.77	17.80[1]	22.84[1]	30.44[1]
21	Op	Co	+	-	1.95	0.12	13.01	28.85[1]	23.76[1]	4.18
23	Co	ES	+	-	0.24	3.26	1.95	30.19[1]	18.45[1]	27.93[1]
24	Co	Op	+	-	2.64	5.48	6.37	4.38	5.26	3.33
27	Op	ES	+	-	8.05	0.03	1.73	0.38	0.05	3.76
28	Ag	Co	+	-	1.40	0.01	3.49	0.24	0.12	2.59
31	Ag	ES	+	-	5.13	4.31	43.30[1]	24.02[1]	14.39[1]	12.38
32	Ag	Co	+	+	6.33	1.25	31.14[1]	51.42[1]	38.60[1]	10.79
40	Op	Co	+	-	7.47	0.22	8.83	0.15	3.91	0.29
41	ES	Co	+	-	11.51					
42	Op	Ag	+	-	10.72	0.85	21.15[1]	6.82	11.46	1.34
47	ES	Op	+	-	10.50	0.50	6.83	1.68	1.37	0.19
48	Co	ES	+	-	6.01	4.61	5.20	14.06[1]	17.74[1]	16.43[1]
50	Co	Op	+	-	0.94	5.43	0.03	9.10	11.15	8.36
51	Co	Ag	+	-	9.63					
53	Op	Ag	+	-	6.85					
54	Op	Co	+	-	6.78	1.16	55.32[1]	112.78[1]	73.75[1]	120.91[1]
56	ES	Ag	+	+	4.07	5.57	19.17[1]	69.39[1]	53.26[1]	14.81[1]
57	Ag	Op	+	-	6.40					
58	Co	Ag	+	+	2.34	0.00	16.84[1]	0.13	4.14	1.89
59	Ag	ES	+	-	3.48	9.22	0.66	10.18	19.25[1]	19.58[1]
61	Ag	Op	+	+	0.98	2.31	0.93	15.91[1]	13.34	7.30
64	Ag	Op	+	-	12.37	1.38	0.07	0.31	0.13	6.32
65	Co	Op	+	-	1.22	3.73	7.93	3.75	5.12	0.06
66	Op	ES	+	+	1.05	4.38	14.75[1]	40.34[1]	38.89[1]	14.40[1]
67	Op	ES	+	-	1.16	0.65	8.78	7.71	4.93	0.17
68	ES	Op	+	+	6.70	5.24	32.03[1]	20.39[1]	32.12[1]	0.06
69	Ag	ES	+	+	1.67	0.56	35.05[1]	18.38[1]	13.35	8.62
71	Op	Ag	+	-	4.79					
72	Ag	ES	+	-	2.76	1.45	22.95[1]	13.44	10.68	2.26
73	Co	Ag	+	-	2.18	0.12	5.85	3.72	4.01	2.50
76	Ag	Co	+	-	0.33	6.73	0.75	4.65	0.90	0.19
77	Co	ES	+	+	27.55[1]	5.64	4.09	0.01	0.05	1.29
79	Co	ES	+	-	0.99	14.05[1]	20.06[1]	12.81	9.11	8.09

Note. ES = emotional Stability; Ag = agreeableness; Op = openness; Co = conscientiousness.
[1] Significant at $\alpha = 2.07 \times 10^{-4}$.

Forced-choice questionnaire. A third booklet consisted of 98 FC bidimensional blocks. Out of them, 79 were made up from items from the GS questionnaire (except for 13 pairs, which contained a direct item from the GS booklets, paired with an inverse item not included in that instrument). There were also sixteen additional blocks made up by items from a different application, and three directed blocks (at positions 25, 43, and 76) to control for response quality. Table 1 summarizes the frequency distribution of the FC blocks by pair of traits. Out of the 79 blocks with items from the GS questionnaire, 24 were formed by two direct items (homopolar blocks); the remaining 55 were heteropolar, consisting of a direct and an inverse item, being the direct one always in the first position. An example of a homopolar block tapping emotional stability and extraversion would be as follows: "Choose the item in each block that is most like you: 'Seldom feel blue/Make friends easily'." Both items have been selected from the Big-Five IPIP Inventory item pool (Goldberg, 1999).

Participants

Seven hundred and five undergraduate students (79.57% female, 20.00% male, and 0.43% missing; age mean and standard deviation, 20.05 and 3.33 respectively), from the first and third courses in the Faculty of Psychology of the Autonomous University of Madrid, answered the GS questionnaire on optical mark reader-ready response sheets. Arguably, this convenience sample might not be the most adequate for a personnel selection context. However, as commented later (see Discussion), a comparison between the GS item and block parameters is more straightforward in a student sample, in which the role of impression management is expected to be weak. Therefore, we deemed appropriate using this dataset for our purposes.

Eight participants were dropped due to having too many missing responses (more than 68), and two more because of failing the directed items (more than one error). Of the remaining 695, 396 (80.36% female, 19.13% male, and 0.51% missing; age mean and standard deviation, 20.86 and 3.21 respectively) also responded to the FCQ on another optical mark reader-ready sheet. No participants were dropped due to missing responses (only 12 vectors had just one missing response), but four were deleted due to failing one or more directed blocks, leaving 392 valid participants. There is a noticeable reduction (313) from the initial sample size (705) to the final one (392). Out of these 313, most of them (299) are missing by design cases, produced because some of the first sample participants were not assessed with the two specific questionnaires required for the current study.

Data analysis

The questionnaires were analyzed with a multidimensional IRT model using the robust maximum likelihood (MLR) method (Yuan & Bentler, 2000) for fitting the item and block responses altogether. The 64-bit Mplus 7.0 software (Muthén & Muthén, 1998-2012) for Windows was used for all analyses. The MplusAutomation 0.7-1 package (Hallquist & Wiley, 2018) for 64-bit R 3.4.3 (R Core Team, 2017) was used to automate some of the analysis procedures.

We tried to fit a model with independent uniquenesses and all the Big-Five traits initially. However, the full-dimensional model had convergence issues with extraversion. Therefore, the items tapping extraversion and the blocks containing an extraversion item had to be dropped. The responses to the remaining 47 blocks and the 86 GS items included in those were finally fitted to a model with the remaining four dimensions. The empirical reliabilities (Equation 20.21; Brown, 2018) of the emotional stability, openness, agreeableness, and consciousness scores were .89, .84, .79 and .82, and .85, .64, .65 and .78, for the GS and FC formats, respectively.

A constrained model was fit for each possible restriction given by the invariance assumption: equal discriminations for a block and each of its corresponding GS items, and a constraint on the block intercept and item parameters given by Equation 3 (using the four possible values of k'). This would result in six contrasts per block, making a total of 282 constrained models. However, given that the GS-item parameters were not available for 8 items (out of the 13 taken from a previous application as explained above, after excluding five of them measuring extraversion), only the first discrimination parameter of the corresponding blocks could be tested for invariance, and therefore 242 constrained models were estimated.

For each of the constrained models, a likelihood ratio test against the unrestricted model was performed as follows: a strictly positive $\chi^2_{S\text{-}B}$ statistic (Satorra & Bentler, 2010) was first computed using the procedure explained by Asparouhov and Muthén (2010). Using a confidence level of .05, the Bonferroni correction for multiple comparisons was applied to these tests, giving a value of $\alpha = .05/242 = 2.07\times10^{-4}$. The parameters for which the p-value of the likelihood ratio test was less than were considered non-invariant.

Results

The correlations of the block parameter estimates with their item counterparts are given in Table 2, along with the descriptive statistics of the deviations (Correlations through MRE columns). Mean error column is the mean difference of the block parameter estimates concerning the GS format estimates (the expected value in the corresponding block, as a linear combination of the item parameters, in the case of the intercept parameters). The MRE column shows the "mean relative error", which is the mean error of the estimates of the blocks relative to the GS items. Negative values, as in these cases, imply a general underestimation in the absolute value of the parameters in the FC format.

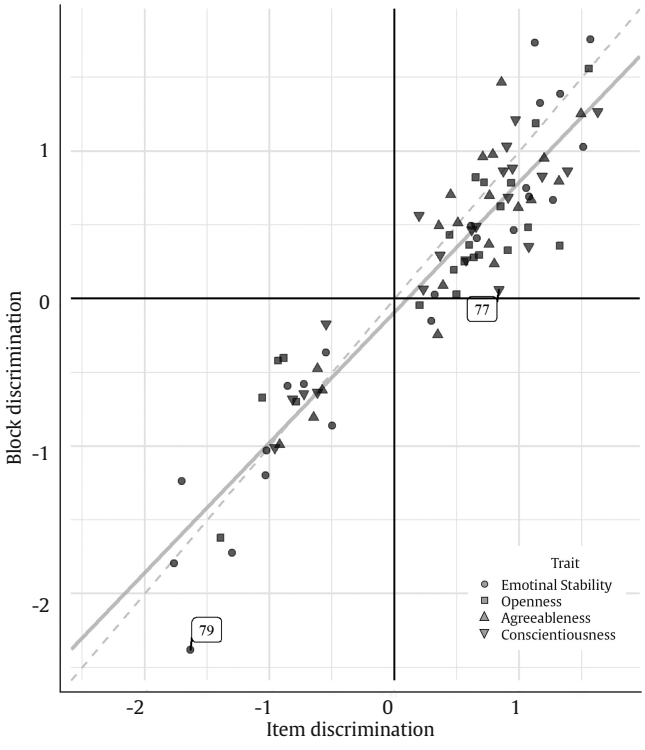

Figure 1. Scatter Plot of the FC-block Discrimination Parameter Estimates against the corresponding GS-item Estimates.
Note. The linear regression trend is shown in continuous light grey. Non-invariant discrimination parameters are annotated with the block code.

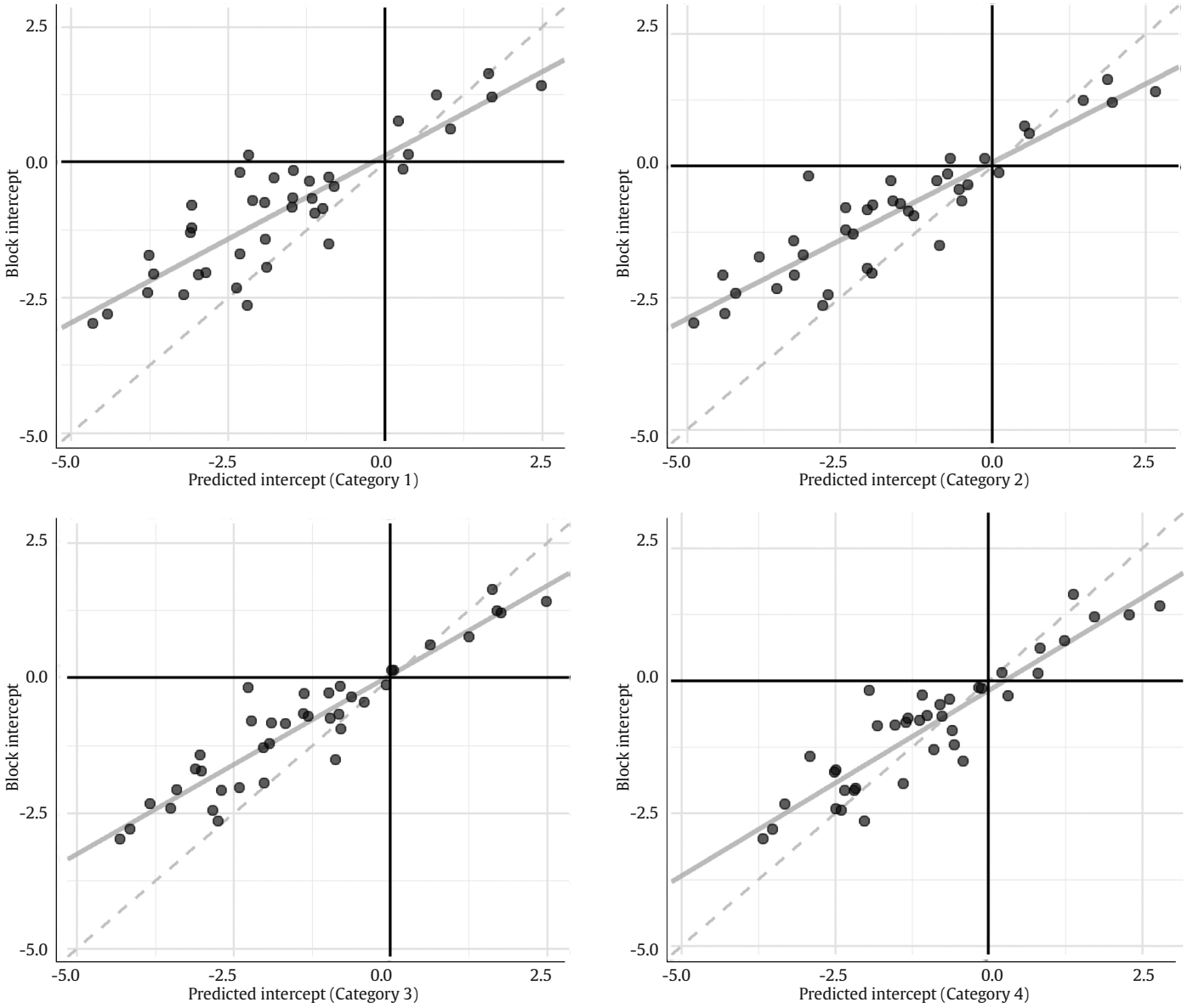

Figure 2. Scatter Plot of the Block Intercept Estimates, against their Values Predicted from the Item Parameters.
Note. The linear regression is shown in continuous light grey.

The last two columns in Table 2 show a summary of the invariance tests. Count column is the absolute frequency of parameters for which the null hypothesis of invariance was rejected. Column % shows the corresponding percentage, relative to the number of parameters of each type. The results of the invariance tests can be seen in detail in Table 3.

Discrimination Parameters

The correlation between both formats was .93, indicating a high correspondence between them. The mean error and mean relative error were negative, implying a slightly negative bias and a general underestimation of the parameters, respectively. That is, there was a slight shrinkage of the parameters towards zero in the FC format. These effects can be appreciated in Figure 1: the regression line intersects the vertical axis slightly below zero and is a bit closer to the horizontal axis than the bisector, which would be the expected regression line in the absence of any type of bias. In the lower right quadrant, we can also see that three of the items reversed their sign when paired in an FC block. Their values in the GS items were already very low though (they were not significantly different from 0), so this was likely due to estimation error.

Despite the deviations from the item estimates, the discrimination parameters were largely invariant. Only two of the null hypotheses of invariance (out of 86) were rejected. These non-invariant parameters are in the last two blocks analyzed, one in the first position and the other in the second position. These results provide strong evidence that the discrimination parameters are invariant between the GS and FC formats.

Intercept Parameters

The correlations of the intercept estimates with their predicted values from the items were also very high in all the cases: all of them were above .90 except with the predictions using the first threshold category. The third threshold category yielded the highest correlation

with the block intercept estimates. The (consistently) lower the mean error, the higher the threshold category, but always positive, in contrast to the discrimination parameters. The mean relative error was negative for all thresholds, manifesting a generalized underestimation in absolute value, similar to that in the discrimination parameters. For the intercept estimates, the third category yielded the lowest mean relative error, followed by the fourth one.

Figure 2 shows how intercept parameter estimates resembled the values predicted from the GS format items. These scatter plots show the tendency of the intercept estimates to be shrunk towards 0 with respect to their predicted counterparts from the items. Also, we can clearly see that the block intercept estimates were better predicted by the third and fourth threshold categories, as seen in Table 2 as well.

The intercept parameters were non-invariant concerning their values predicted from the GS format estimates in 10 to 15 cases, depending on the item threshold category considered. The fourth one had the lowest number of non-invariant parameters, followed by the third one with 12. The second one had the highest number. The intercept estimate was invariant for all the threshold categories in 17 out of the 39 blocks for which the intercept parameter could be predicted (43.59%). Only in three of them, the intercept parameter was found to be non-invariant for all the categories. The rest of the blocks had non-invariant intercept parameters in one to three threshold categories.

Discussion

From the results above, we can conclude that the FC format generally satisfies the invariance assumption of the MUPP-2PL model. Apart from the high rates of invariance, we found high correlations between the parameters of the FC and GS formats, although there seemed to be a general trend of the FC format estimates to be lower in absolute value.

Some of the parameters failed to pass the invariance test, yielding evidence of violations of the invariance assumption. The intercept parameters were the most affected, whereas only two discrimination parameters were non-invariant. Due to this low rate of non-invariance, hypothesizing about causal phenomena would be highly speculative.

The intercept parameters showed some recognizable patterns of non-invariance. Figure 3 plots the deviation of the non-invariant intercept estimates concerning their predicted values from the GS-item estimates. This figure shows that most of the intercept parameters had a positive deviation regardless of the threshold category. The fourth category was an exception, as there was an equal number of positive and negative errors among the non-invariant parameters.

Only a few estimates deviated from their predictions from the GS format consistently. Some properties of the items seem to be affecting the invariance of the intercept parameters. For example, emotional stability and openness seemed to be more involved in the non-invariant intercept parameters. Also, there seemed to be an association between deviation direction and block polarity for this threshold category, as most of the negative errors were in homopolar blocks (i.e., with a direct item in the second position), while most of the positive errors were in heteropolar blocks. Moreover, violations of invariance were more prevalent with emotional stability items in the second position and openness items in the first one, suggesting a complex interaction effect among the two latent traits, the item position within the block, and the item and block polarities. Morillo (2018) provides an extensive discussion around these violations and the factors that likely induce such a lack of invariance in the parameters.

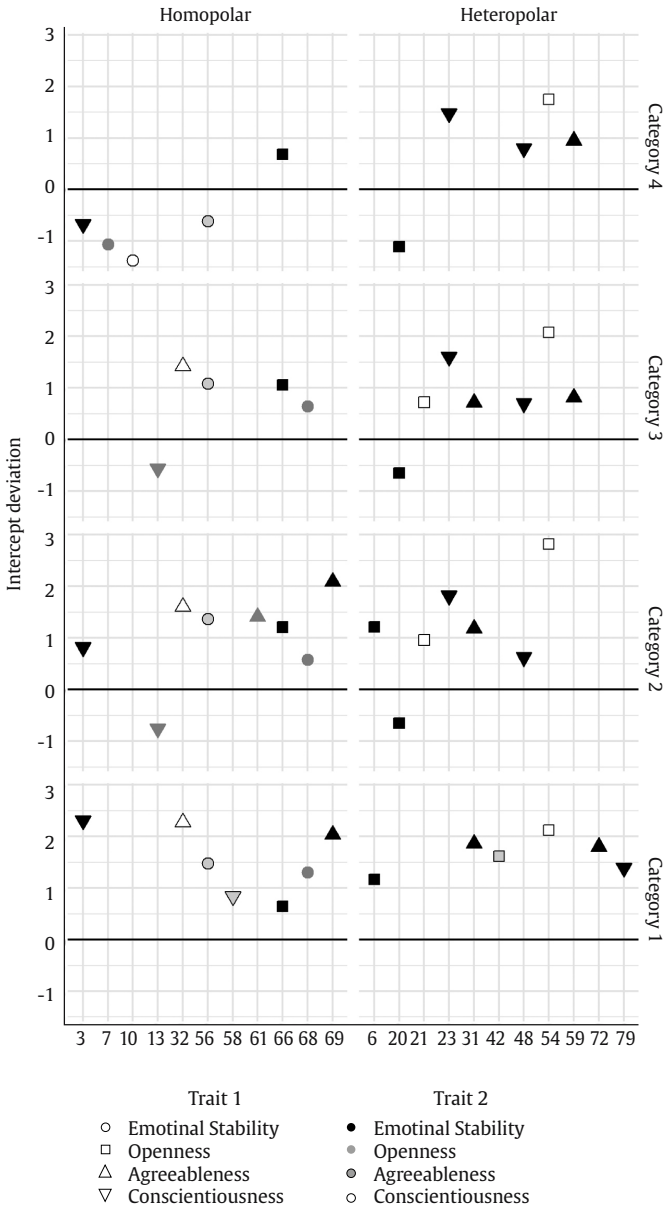

Figure 3. Deviation of the Non-invariant Block Intercept Parameters with respect to their Predicted Values from the Item Parameters.

Implications for the Practice of Personnel Selection

The fact that the invariance between the GS and the FC formats can be safely assumed has a great practical relevance: it enables the practitioner to safely build multidimensional FC instruments based on the parameter estimates of the individual items. The designer only should be careful to avoid certain pairings that could lead to violations of invariance, as these would likely reduce the validity of the measures. A good starting point is the recommendations by Lin & Brown (2017): balancing item desirability indices and avoiding pairing items with shared content and/or conceptually-similar latent constructs. However, these recommendations require the items to be calibrated on a social desirability scale, and their contents to be submitted to a qualitative analysis. Also, we believe that further research, probably in experimental settings, would help identify other conditions that may produce non-invariant parameters.

The process of constructing a multidimensional FCQ is thus not as straightforward as simply pairing items tapping different latent traits.

Nevertheless, the practitioner can rely upon GS estimates of the item parameters to assess a priori the potential validity of the new instrument. A procedure of FCQ construction based on this principle could be outlined as follows: (1) to calibrate a set of items in a GS format (or use the estimates from a previously calibrated instrument); (2) to decide on certain design criteria (e.g., balance of trait and polarity pairings, pairing exclusions based on expected violations of the invariance assumption, etc.); (3) to pair the items in FC blocks attending to such criteria; (4) to apply the FC instrument to an equivalent sample; and (5) to calibrate the FCQ on the new sample data and obtain the latent trait scores. If properly designed, the new FC instrument should have parameters comparable to the original items and thus similar validity. Note however that this would not allow applying the method outlined here for testing the invariance assumption. For that to be possible, the newly created FCQ would need to be calibrated with the same sample as the GS items; this will not be generally possible in an applied context. Nevertheless, the parameter correspondence could be examined using multivariate descriptive methods.

Limitations

The research design used in this manuscript had some issues that did not allow us to accurately separate the effects of the latent trait, polarity, and item position within the block. Nevertheless, taking into account the possible violations of the invariance assumption should be paramount for research purposes. Further studies should aim to overcome two limitations: (1) to design FCQs that balance the order of the inverse item in heteropolar blocks and (2) to calibrate the parameters of the whole set of items in both formats. Using a different response format for the items could also be advantageous, such as an even number of GS response categories, or a dichotomous format. More complete response vectors would also be desirable, as the present one lacked a large number of responses for the FC blocks in comparison with the items.

This study has some other limitations worth highlighting. It is especially worth pointing out the problems found when estimating the models with the "extraversion" trait. We could not find convergence due to the latent correlation matrix becoming singular, as the correlations between the dimension of extraversion and the others approached 1. This fact may suggest some property of the multidimensional FC format affecting specifically this trait. Whatever the actual explanation is, it should not be overlooked if we want the results to be fully extrapolated to the Big Five model, and to other theoretical models the FC format may be applied to.

The use of the response dataset may also be criticized, as it had been obtained from a sample of students. The reader should also note that the invariance assumption was tested under a situation of honest responding or "direct-take". Of course, a high-stakes situation could imply stronger violations of invariance than the direct-take one. The fulfillment of the invariance assumption in an honest test-taking context is a necessary condition, as the process of pairing stimuli must be validated beforehand. However, this condition is not sufficient for an impression management context. In a high-stakes context, other factors accounting for the impression management attempts may emerge, adding further complexity to the measure and its treatment. Further studies applying the methodology outlined here will allow generalizing these results to actual personnel selection settings.

Finally, the application of the questionnaire to a calibration sample should provide evidence that the response data to a multidimensional FCQ are valid. Although the invariance assumption discussed here is a necessary condition, it does not guarantee the validity of the FCQ latent trait scores. This issue has not been investigated for the MUPP-2PL model, but there is evidence of latent score validity in other FC formats and models (Lee, Joo, Stark, & Chernyshenko, 2018; Lee, Lee, & Stark, 2018).

Conclusions

This study introduces a methodology that allows testing the assumptions of the MUPP-2PL model for paired FC blocks. The application of this method may open up further research lines, as the previous discussion suggests. More importantly, we have provided evidence that the invariance assumption between the GS and the FC formats holds to a large extent. This finding provides the practitioner with tools and criteria to seriously consider the design of multidimensional FC instruments to measure personality and other non-cognitive traits of high importance in work-related assessment settings. Particularly, our results have practical relevance for building multidimensional FCQs by using previously calibrated items. Evidence of the invariance assumption legitimates the design of FC blocks using the already known parameters of the items as a proxy for the block parameters. Given the large number of applications of personality questionnaires in GS formats, this obviously implies a considerable cost reduction.

We have outlined a general procedure based on those principles, giving some guidelines to mitigate possible violations of the invariance assumption. Also, assuming invariance may allow using the GS format estimates to optimize certain design criteria for FCQs. Such criteria may even be implemented in automatic assembling algorithms (Kreitchmann, Morillo, Ponsoda, & Leenen, 2017; Yousfi & Brown, 2014), making the design of FCQs more efficient and cost-effective.

Conflict of Interest

The authors of this article declare no conflict of interest.

References

Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement, 13*, 193-216. https://doi.org/10.1177/014662168901300211

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19,* 269-290. https://doi.org/10.1177/014662169501900306

Asparouhov, T., & Muthén, B. O. (2010). *Computing the strictly positive Satorra-Bentler chi-square test in Mplus.* Retrieved from https://www.statmodel.com/examples/webnotes/SB5.pdf

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9-30. https://doi.org/10.1111/1468-2389.00160

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational & Organizational Psychology, 69*, 25-39. https://doi.org/10.1111/j.2044-8325.1996.tb00597.x

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 technical manual.* Surrey, UK: SHL.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Pub. Co.

Brown, A. (2018). Item Response Theory approaches to test scoring and evaluating the score accuracy. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley handbook of psychometric testing.* London, UK: John Wiley & Sons. https://doi.org/10.1002/9781118489772.ch20

Brown, A., & Bartram, D. (2011). *OPQ32r technical manual.* Surrey, UK: SHL.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71,* 460-502. https://doi.org/10.1177/0013164410375112

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18,* 267-307. https://doi.org/10.1207/s15327043hup1803_4

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational & Organizational Psychology, 67,* 89-100. https://doi.org/10.1111/j.2044-8325.1994.tb00553.x

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London.*

Series A, Mathematical or Physical Character, 222(594-604), 309-368. https://doi.org/10.1098/rsta.1922.0009
Ghiselli, E. E. (1954). The forced-choice technique in self-description. *Personnel Psychology, 7,* 201-208. https://doi.org/10.1111/j.1744-6570.1954.tb01593.x
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.
Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 0*, 1-18. https://doi.org/10.1080/10705511.2017.1402334
Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167-184. https://doi.org/10.1037/h0029780
Judge, T. A., Rodel, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*, 875-925. https://doi.org/10.1037/a0033901
Kreitchmann, R. S., Morillo, D., Ponsoda, V., & Leenen, I. (2017). *An optimization procedure for assembling multidimensional forced-choice blocks.* International Meeting of the Psychometric Society 2017. Zürich, Switzerland. Retrieved from https://www.psychometricsociety.org/sites/default/files/IMPS_2017_Talks_w_Cover.pdf
Lee, P., Joo, S.-H., Stark, S., & Chernyshenko, O. S. (2018). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement* (Advance online publication). https://doi.org/10.1177/0146621618768294
Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences, 123,* 229-235. https://doi.org/10.1016/j.paid.2017.11.031
Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*, 389-414. https://doi.org/10.1177/0013164416646162
Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*, 73-79. https://doi.org/10.1027/1614-2241.4.2.73
Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008
McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x
Morillo, D. (2018). *Item response theory models for forced-choice questionnaires* (Doctoral dissertation). Universidad Autónoma de Madrid. Retrieved from https://repositorio.uam.es/handle/10486/686097
Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., De la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 40*, 500-516. https://doi.org/10.1177/0146621616662226
Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65*, 693-717. https://doi.org/10.1146/annurev-psych-010213-115134
Salgado, J. F. (2017). Moderator effects of job complexity on the validity of forced-choice personality inventories for predicting job performance. *Journal of Work and Organizational Psychology, 33,* 229-238. https://doi.org/10.1016/j.rpto.2017.07.001
Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88,* 797-834. https://doi.org/10.1111/joop.12098
Salgado, J. F., & Lado, M. (2018). Faking resistance of a quasi-ipsative forced-choice personality inventory without algebraic dependence. *Journal of Work and Organizational Psychology, 34,* 213-216. https://doi.org/10.5093/jwop2018a23
Salgado, J. F., & Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*, 3-30. https://doi.org/10.1080/1359432X.2012.716198
Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series, 1968*(1). https://doi.org/10.1002/j.2333-8504.1968.tb00153.x
Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75,* 243-248. https://doi.org/10.1007/s11336-009-9135-y
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. https://doi.org/10.1111/j.2044-8325.1991.tb00556.x
Scott, W. A. (1968). Comparative validities of forced-choice and single-stimulus tests. *Psychological Bulletin, 70*, 231-244. https://doi.org/10.1037/h0026262
Sisson, E. D. (1948). Forced choice - The new Army rating. *Personnel Psychology, 1*, 365-381. https://doi.org/10.1111/j.1744-6570.1948.tb01316.x
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. https://doi.org/10.1177/0146621604273988
Tett, R., Rothstein, M. G., & Jackson, D. J. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742. https://doi.org/10.1111/j.1744-6570.1991.tb00696.x
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273-286. https://doi.org/10.1037/h0070288
Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*, 60-62. https://doi.org/10.1214/aoms/1177732360
Yousfi, S., & Brown, A. (2014). *Optimal forced-choice measurement for workplace assessments.* The 9th Conference of the ITC: Global and Local Challenges for Best Practices in Assessment. San Sebastián, Spain.
Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 3,* 165-200. https://doi.org/10.1111/0081-1750.00078