*Papeles del Psicólogo / Psychologist Papers, 2022 Vol. 43(1), pp. 29-35*
https://doi.org/10.23923/pap.psicol.2982
http://www.papelesdelpsicologo.es
http://www.psychologistpapers.com

# Building adaptive forced-choice tests "On The Fly" for personality measurement

**Abad, Francisco J.[1]; Kreitchmann, Rodrigo S.[1]; Sorrel, Miguel A.[1]; Nájera, Pablo[1]; García-Garzón, Eduardo[2]; Garrido, Luis Eduardo[3] and Jiménez, Marcos[1]**
[1]*Universidad Autónoma de Madrid, España.* [2]*Universidad Camilo José Cela, España.* [3]*Pontificia Universidad Católica Madre y Maestra, República Dominicana*

*Los nuevos desarrollos metodológicos y tecnológicos de la última década permiten resolver, o al menos atenuar, los problemas psicométricos de los test de elección forzosa (EF) para la medición de la personalidad. En estas pruebas, a la persona evaluada se le muestran bloques de dos o más frases de parecida deseabilidad social, entre las que debe elegir aquella que le represente mejor. De esta manera, los test de EF buscan reducir los sesgos de respuesta en pruebas de autoinforme. No obstante, su uso no está exento de riesgos y complicaciones si no se elaboran adecuadamente. Afortunadamente, los nuevos modelos psicométricos permiten modelar las respuestas en este tipo de test, así como optimizar su construcción. Más aún, permiten la construcción de Test Adaptativos Informatizados de EF (TAI-EF) "on-the-fly", en los que cada bloque se construye en el mismo momento de aplicación, emparejando óptimamente las frases de un banco previamente calibrado.*

***Palabras clave:*** *Personalidad, Elección forzosa, Test adaptativo informatizado, On-the-fly.*

*The new methodological and technological developments of the last decade make it possible to resolve or, at least, attenuate the psychometric problems of forced-choice (FC) tests for the measurement of personality. In these tests, the person being tested is shown blocks of two or more sentences of similar social desirability, from which he or she must choose which one best represents him or her. Thus, FC tests aim to reduce response bias in self-report questionnaires. However, their use is not without risks and complications if they are not created properly. Fortunately, new psychometric models make it possible to model responses in this type of test and to optimize their construction. Moreover, they allow the construction of "on the fly" computerized adaptive FC tests (CAT-FC), in which each item is constructed on the spot, optimally matching sentences from a previously calibrated bank.*

***Key words:*** *Personality, Forced-choice, Computerized adaptive tests, On-the-fly.*

Several meta-analysis studies support the predictive role of personality variables, measured by self-report, in organizational, educational, and health contexts (e.g., Barrick & Mount, 1991; Judge et al., 2013; Otero et al., 2020). We know, for example, that Conscientiousness and Emotional Stability have generalized predictive validity across different occupations and criteria, while other dimensions, such as Openness to Experience, Agreeableness, or Extroversion, are also relevant in particular contexts and in the prediction of specific criteria. In the educational domain, Conscientiousness and Emotional Stability play an important role in predicting academic performance (Poropat, 2009; Richardson et al., 2012), whereas Conscientiousness and Agreeableness predict counterproductive behaviors (e.g., cheating on exams; Cuadrado et al., 2021).

Despite these results, doubts about the sensitivity of self-report to the effects of social desirability and faking have accompanied these tests since their inception. Specifically, in selection contexts, it is expected that candidates will distort their answers to give a more positive image of themselves, either by self-deception or deliberately in order to be chosen. These distortions produce strong increases in mean scores in the perceived desirable direction and reduce the reliability and variability of scores (Viswesvaran & Ones, 1999; Salgado, 2016). These effects are found both in experimental studies, when comparing honest and dishonesty-induced responses, and—although more attenuated—in applied contexts, when comparing samples of job applicants and incumbents (Salgado, 2005). Therefore, without prejudice to the fact that positive results on the predictive validity of personality scores are found in real contexts, previous evidence suggests that the ranking of candidates could be different depending on whether they fake their answers or not. Salgado (2005) describes some strategies to reduce the effects of faking, such as informing test takers that there is a possibility of being penalized if they distort their answers or using specific scales made from samples of applicants.

Another possibility is to use response formats that are more robust to faking. In the field of self-report tests, a distinction can be made between the traditional Likert scale format (i.e., items or statements for which the respondent has to indicate his/her degree of agreement) and the forced-choice format (i.e., blocks of one or more statements, from which the respondent must make a choice (e.g., indicate the one that best represents him/her) or a ranking (e.g., rank the items in order, partially or completely, according to the degree to which they describe him/her). Table 1 shows examples of forced-choice items with different formats.

The Likert format is susceptible not only to the effects of social desirability or faking, but also to the presence of other response biases such as acquiescence bias, negativity bias, central tendency, or

*Received: 1 November 2021 - Accepted: 16 December 2021*
*Correspondence:* Abad, Francisco J.
*E-Mail: fjose.abad@uam.es.*

............
*This work was carried out with financial support from the Ministerio de Ciencia, Innovación y Universidades (Proyecto PSI2017-85022-P) and by the Cátedra de Modelos y Aplicaciones Psicométricos (Instituto de Ingeniería del Conocimiento and Universidad Autónoma de Madrid).*

extreme response bias, etc. The existence of response biases can distort the factor structure of the scale and lead to maladjustment (e.g., Abad et al., 2018), as well as produce an overestimation of reliability and alter convergent validity estimates. In contrast, these biases do not apply to the forced-choice format. Specifically, it is expected that, if blocks are formed with items matched in social desirability, the susceptibility to misrepresentation is reduced.

## PROBLEMS OF SCORES OBTAINED IN FORCED-CHOICE TESTS

Despite the above, the use of the forced-choice format has not been free of controversy. First, its greater resistance to faking has been questioned (e.g., Heggestad et al., 2006). However, more recent meta-analysis studies suggest that the effect of faking is smaller in forced-choice tests (Cao & Drasgow, 2019; Martínez & Salgado, 2021). Second, forced-choice tests can result in scores with *ipsative* properties, in which the interpretation of a score is relative to the rest of the scores belonging to the same subject. For example, a highly organized and sociable person may have the same response as a person who is not very organized or sociable, because they both consider themselves to be *more* organized than sociable. If the scores are completely ipsative, the sum of the scores of each subject will result in the same constant value and *normative* interpretations (e.g., concluding that the first person is more organized than the second) would be risky. In these cases, the application of traditional psychometric analysis techniques will result in methodological artifacts (Hicks, 1970). For example, the average expected correlation between the dimensions will tend to be negative. Similarly, the correlations between scores on those dimensions and any external criterion will be zero. These results come from the negative covariances that occur when forcing a person to choose one statement over another. For example, consider the extreme case in which a test includes 20 blocks of two items, one scoring positively on Extroversion and the other scoring positively on Conscientiousness. If we add +1 in Extroversion for each Extroversion item chosen and +1 in Conscientiousness for each Conscientiousness item chosen, the correlation between the two scales will be -1 and the sum of the scores on both scales will be 20, regardless of the choices of the test takers.

Ipsativity is not an all-or-nothing question, nor is it associated with the format itself, but rather it depends on the design of the test and the blocks (e.g., number of items per block, unidimensional/multidimensional nature of the blocks, direct/inverse polarity of the items forming the blocks, number of dimensions assessed, correlation between the dimensions measured, and scoring mode). In this sense, scores on a forced-choice test can be any of the following (Hicks, 1970): (a) fully ipsative; (b) quasi- or partially ipsative; or (c) normative. Normative scores can be obtained, for example, if items in the same block belong to the same dimension. Scores can become partially ipsative if, for example, respondents partially—rather than completely—order the alternatives, the scales differ in the number of items, or one of the dimensions is not scored. Quasi-ipsative scales result in scores that do not sum to a constant for all individuals, but may maintain some interdependence, i.e., the problem is reduced, but may not be eliminated (Brown & Maydeu-Olivares, 2018). Some meta-analyses show that quasi-ipsative tests have higher predictive validity (Salgado et al., 2015; Salgado & Táuriz, 2014) and are more robust to faking (Martínez & Salgado, 2021)

The four most frequent formats of forced-choice tests are (see Table 1): (a) choosing the item that best describes you from two statements (PICK-PAIR), (b) choosing the item that best describes you from more than two statements (PICK), (c) choosing the item that best describes you and the item that least describes you (MOLE, from "MOst and LEast"), and (d) ranking the options according to the degree to which they describe you (RANK). As for the traditional scoring, in the PICK

**TABLE 1**
**EXAMPLES OF FORCED-CHOICE FORMAT ITEMS**

| Type of format (in parentheses, for each sentence, measured dimension and polarity) | Choice/score | Choice /score |
|---|---|---|
| PICK.PAIR. *Choose the phrase that best represents you:*<br>A. I believe that others have good intentions (Ag+)<br>B. I make lists of things to do (Co+) | A /<br>Ag: +1<br>Co: +0 | B/<br>Ag: +0<br>Co: +1 |
| PICK. *Choose the phrase that best represents you:*<br>A. I feel relaxed most of the time (Es+)<br>B. I believe that the others have good intentions (Ag+)<br>C. I make lists of things to do (Co+) | A/<br>Es: +1<br>Ag: +0<br>Co: +0 | B/<br>Es: +0<br>Ag: +1<br>Co: +0 |
| MOLE (e.g., Heggestad et al., 2006). *Indicate the sentences that represent you best (↑) and worst (↓):*<br>A. I avoid difficult reading material (Op-).<br>B. I only feel comfortable with friends (Ex-)<br>C. I believe that others have good intentions (Ag+)<br>D. I make lists of things to do (Co+) | C↑; B↓/<br>Op: +0<br>Ex: -(-1)<br>Ag: (+1)<br>Co: +0 | B↑; C↓/<br>Op: +0<br>Ex: -(+1)<br>Ag: (-1)<br>Co: +0 |
| *RANK. Rank the sentences according to the degree to which they represent you, from "most like you" to "least like you":*<br>A. I feel relaxed most of the time (Es+)<br>B. I believe that others have good intentions (Ag+)<br>C. I make lists of things to do (Co+)<br>D. I like to learn new things (Op+) | A>B>C>D/<br>Es: +4<br>Ag: +3<br>Co: +2<br>Op: +1 | D>B>C>A/<br>Es: +1<br>Ag: +3<br>Co: +2<br>Op: +4 |

Note: Ag: Agreeableness; Op: Openness; Es: Emotional stability; Ex: Extroversion; Co: Conscientiousness; +: direct item or positive polarity; -: inverse item or negative polarity and PICK-PAIR formats a score of +1 can be given on the dimension if the polarity of the chosen item is positive (see Table 1) or -1, if it is negative (i.e., reverse item). In the RANK format, values between 1 and *K* can be assigned, *K* being the number of sentences to be ranked in order, whereas in the MOLE format, scores -1, 0, or 1 can be assigned, depending on the specific choice and the polarity of the selected items (see two examples in Table 1). Hontangas et al. (2015, 2016) found, by simulation, that the MOLE format provided similar results to RANK, and both cases were superior to PICK. However, Cao and Drasgow (2019) found the PICK format to be more resistant to faking than the MOLE format, indicating that the latter, in addition, involves a higher cognitive load in responding.

## SCORING FORCED-CHOICE BLOCKS BASED ON IRT

In recent years, it has been suggested that many of the problems of forced-choice test scores may be due to the classical scoring procedure itself and can be overcome by modeling the responses from item response theory (IRT; e.g., Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015, 2016; Morillo et al., 2016). IRT allows us to model the probabilities of the response to blocks as a function of trait levels, which makes it possible to achieve, under certain conditions, a normative interpretation of the scores (i.e., it enables comparisons between individuals). The use of IRT models has several advantages (Olea et al., 2010): (a) it allows us to assess the accuracy for each trait level, instead of assuming that all individuals are assessed with the same reliability; (b) it allows us to obtain scores on the same scale, even when different items are applied; and (c) it allows us to develop advanced applications, such as computerized adaptive tests (CATs). The main characteristic of CATs is that the items administered are adjusted to the level of trait that the person being evaluated manifests progressively, according to his or her responses to previous items. The use of a CAT makes it possible to obtain more efficient measures (same level of precision in less time), as well as measures with a more homogeneous level of precision across the trait level. Several types of IRT models have been proposed to describe the item comparison process within a block, among which MUPP (multi-unidimensional pairwise preference) and TIRT (Thurstonian item response theory) stand out.

The MUPP model was developed by Stark et al. (2005) for blocks of two items, each measuring a different dimension. First, a model is defined for the probability that a person agrees with the content of an item. This probability can be assumed to follow either a dominance model (Morillo et al., 2016) or an ideal point model (Stark et al., 2005). In a dominance model the probability of agreement with an item (e.g., "I believe others have good intentions") increases as a function of the trait level (e.g., Agreeableness). In an ideal point model, the response probability function is unimodal; that is, the probability of agreement increases as a function of trait level until it reaches a maximum and then decreases. For example, the probability of agreement with the item "Sometimes I can persuade my friends to do things my way" may be maximum for people who have some capacity of persuasiveness, but lower for people who never persuade their friends or for people who always persuade their friends. Second, from these item agreement probabilities, the probability of preferring one item over another within a block can be obtained (see, for example, Morillo et al., 2016).

The TIRT model (Brown & Maydeu-Olivares, 2011) is based on Thurstone's law of comparative judgment and assumes a dominance model. In this model, the response to each block is broken down into a set of binary comparisons. For example, suppose someone declares that, from a block of three items, the statement that most represents him or her is B and the one that least represents him or her is A. This ordering (B > C > A) could be represented by three variables, one per binary comparison: $X_{AB} = 0$ (i.e., he/she prefers item B to A), $X_{BC} = 1$ (i.e., he/she prefers item B to C), and $X_{AC} = 0$ (i.e., he/she prefers item C to A). Once these variables have been created, IRT models can be estimated by factor analysis (Brown & Maydeu-Olivares, 2012). In the case of two-statement blocks, Morillo et al. (2016) show that, when the dominance model is assumed, MUPP is equivalent to TIRT.

Which model is better? Ideal point MUPP models are more flexible, but perhaps unnecessarily complex. The decisive question would be whether or not to accept the need to use items with unimodal probability functions. Items such as "Sometimes I can persuade my friends to do things my way" are often discarded in prior psychometric analysis and are often ambiguous and even frustrating to respondents (Brown & Maydeu-Olivares, 2010). Nevertheless, some of the greatest successes in applying IRT to forced-choice items have been achieved with ideal-point models.

In any case, the advantage of using IRT models lies in obtaining scores with lower or even zero ipsativity, although the degree to which this is achieved will depend on the test design, as described below.

## GENERAL FACTORS AFFECTING THE EFFICACY OF FORCED-CHOICE TESTING

As mentioned above, IRT models do not necessarily result in scores with normative properties. Frick et al. (2021) point out some factors that affect the efficacy of forced-choice item construction in the case of dominance models, although their conclusions cannot be considered definitive. First, their most important recommendation is to use both positive homopolar blocks, consisting of items that measure dimensions in the same direction (e.g., A. *I think it is exciting to talk to many different people {Ex+}*; B. *I feel comfortable with myself {Es+}*), and heteropolar blocks, made up of items that measure dimensions in opposite directions (e.g., A. *I like talking to strangers {Ex+}*; B. *I worry about things {Es-}*). For example, using only blocks of the first type may make it more difficult to tell whether someone has chosen item A due to having high extroversion or due to having low stability, whereas including blocks of the second type will help distinguish between the two profiles. Despite this, the need to use heteropolar blocks is debatable, as it may be more difficult to match items on social desirability (Bürkner et al., 2019; Lee & Joo, 2021), thus facilitating faking, which this format is intended to prevent. On the other hand, Morillo et al. (2016) and Kreitchmann et al. (2021) have shown that it is possible to estimate accurately without heteropolar blocks as long as optimal assembly is carried out and there is sufficient range in the weight of the items (which was also pointed out by Frick et al., 2021).

Another important factor is the size of the block. Increasing their size (e.g., using triplets) may reduce ipsativity, but it also increases cognitive load by requiring more comparisons per block (Sass et al., 2020). In fact, Frick et al. (2021) find similar reliability when comparing pairs and triplets if the number of binary comparisons is held constant. Another problem with blocks of more than two items is that, when applying TIRT in the absence of heteropolar blocks, reliability tends to be overestimated.

Other factors relevant to the presence of ipsativity are the correlations between the dimensions and the number of dimensions. The lower the number of dimensions measured or the higher the positive correlation, the higher the ipsativity. For example, from the results in their simulation studies, Bürkner et al. (2019) suggest that with five or fewer dimensions and homopolar blocks accurate measurements cannot be achieved, while accurate results were obtained with 30 dimensions (these authors did not consider intermediate cases, with between 6 and 29 factors). Fisher et al., (2019) are also pessimistic about the use of TIRT, finding worse criterion-referenced validity in selection contexts. Indeed, there are several empirical studies finding that test reliability may be low (e.g., Kreitchmann et al., 2019) or that correlations between traits may be distorted (e.g., Morillo et al., 2016). Presumably, these inconsistencies between studies are due to the difficulty in constructing good forced-choice homopolar blocks.

## THE CONSTRUCTION OF FORCED-CHOICE BLOCKS

The key to success in the construction of a forced-choice test is the matching of items on social desirability, taking into account the information provided by the block as a whole. Regarding the matching of desirability, expert ratings (or ratings of samples similar to the one under evaluation) are usually used to score the social desirability of the items. On this point, Pavlov et al. (2021) underscore the importance of matching items not only for social desirability, but also taking into account the consensus of the judges in the assessment.

Regarding the formation of information blocks, the use of an IRT model makes it possible to anticipate how much information the block will provide when applied (i.e., the degree to which it will reduce the error variance of the estimated trait levels) and to assemble items into blocks to maximize the information. However, the size of the potential universe of blocks is often a problem. For example, assembling 60 items into 30 blocks of 2, results—approximately—in 2.92 X $10^{40}$ possible questionnaires (Kreitchmann et al., 2021). To solve this problem, Kreitchmann et al. (2021) adapt the genetic algorithm NHBSA (node histogram-based sampling algorithm; Tsutsui, 2006) to the problem of assembling items into blocks and provide a user-friendly implementation in Shiny that enables the design of a forced-choice test (https://psychometricmodelling.shinyapps.io/FCoptimization/). Kreitchmann et al. (2021) found that the proposed algorithm was more efficient than the existing methods (e.g., random with content limitations or brute force). In summary, the quality of a forced-choice test depends, as in traditional tests, on the psychometric quality of its components: the blocks of which it is comprised.

## COMPUTERIZED ADAPTIVE PERSONALITY TESTS WITH LIKERT FORMAT

In the personality domain, a number of examples can be found of CATs with Likert scales for measuring the Big Five. Of note is the pioneering work of Reise and Henson (2000) for the NEO-PI-R in which they found that a unidimensional-CAT of only four items per facet (i.e., reducing the length by half) provided accurate recovery of trait levels. CATs based on multidimensional models assuming correlated factors (e.g., Makransky et al., 2013; Nieto et al., 2018) and based on the bifactor model (Nieto et al., 2018), applied within each personality domain (e.g., Extroversion), have also been developed. Multidimensional CATs show some advantage when taking into account the correlations between the different facets (e.g., in the study by Makransky et al., 2013, a high average correlation of 0.7 was obtained for the facets of the emotional stability domain). The studies of Nieto et al. (2017; 2018) investigated the correlations between scores obtained on CATs with those obtained on the full bank. For the domains, with 12 items per domain, average correlations of 0.89 were reached for the unidimensional CAT (and for the short scales), and 0.94 for the multidimensional CATs (Nieto et al., 2018). These, in addition, provided a better balance in the proportion of items applied in each facet (i.e., higher content validity). For the facets, the multidimensional CATs achieved a lower average correlation than the unidimensional ones (0.87 vs. 0.95), but with half the number of items.

## BUILDING ADAPTIVE FORCED-CHOICE CATS

The advantages of a CAT may become especially important in items with few response categories, such as in the PICK-PAIR format, since in these cases the range of trait levels for which the item is accurate is narrow. There are multiple forced-choice CATs (FC-CATs), the most famous being the TAPAS (e.g., Stark et al., 2014), which measures 22 personality dimensions and is understood as an “a la carte test”, it being possible to choose, for example, the dimensions to be assessed, the type of test (adaptive or fixed), and the format (e.g., binary, polytomous, unidimensional forced-choice, or multidimensional forced-choice) depending on the context of application (e.g., greater or lesser prediction of social desirability). Adaptive versions allow the length to be reduced by half (Drasgow et al., 2012).

The effectiveness of a multidimensional FC-CAT depends on: (a) the bank of blocks assembled and (b) the selection rule. Regarding the first point, the information in the previous sections for the construction of fixed tests is applicable. Blocks can be paired, for example, according to a genetic algorithm, producing optimal banks from which to adaptively select blocks. As for the selection rule, there are different variants. In one-dimensional models, the error variance is inversely proportional to the test information, which is the sum of the information functions of the items. Similarly, the error variance-covariance matrix in multidimensional models is the inverse of the information matrix. Despite the apparent similarity, this difference implies that different rules (e.g., T-rule: maximize the information of each dimension when adding the item; A-rule: minimize the error variance of each dimension when adding the item) give different results.

For example, Kreitchmann et al. (submitted for publication) started with a bank of 240 items (48 items per dimension) that they assembled into a bank of 120 blocks. The number of possible blocks, excluding the unidimensional ones, was 23,040. In this case, results were compared for CATs of different length (i.e., 30 and 60 blocks) and selection rule (e.g., T-rule and A-rule), starting with a bank constructed according to the genetic algorithm or a bank made up of random blocks. For the best selection rule, it was found that, on average, the use of an optimal bank versus a random bank could increase the reliability coefficient by 0.05 points (from 0.80 to 0.85) and, more importantly, it could reduce the ipsativity of the scores (the negative bias of the correlations between dimensions and in the relationship with the criterion was reduced by 0.04 points). Regarding the selection rule, Kreitchmann et al. (submitted for publication) found that, consistent with previous research (e.g., Mulder & van der Linden, 2009), the A-rule, by directly minimizing error variances, provided the best results. This result is important, since some researchers use the T-rule for computational efficiency (Chen et al., 2020).

## BUILDING ON-THE-FLY ADAPTIVE FORCED-CHOICE CATS

As mentioned, the use of a genetic algorithm allows the optimization of a fixed test or a bank. The next natural step is to build blocks "on the fly"; that is, to assemble the statements spontaneously into blocks at the time of applying the CAT. This idea is at the heart of TAPAS, which starts with a large set of calibrated items (statements) from which, using the MUPP model, a giant set of potential blocks is derived and from this set the most informative block is selected at any given moment. This procedure is not free of assumptions, since it assumes the veracity of the MUPP model and the absence of context effects (i.e., the performance of each item does not depend on the item with which it is paired). Although there may be context effects, this invariance assumption can be reasonably sustained in practice (Lin & Brown, 2017; Morillo et al., 2019). Lin and Brown (2017) suggest that context effects can be reduced by pairing items on social desirability (otherwise, the item that is most clearly desirable will be chosen more because it is perceived as the "right answer"), and the same authors indicate that the inclusion within the same block of items that are similar in content should be avoided (e.g.,, *I am a lively person in conversation*; I *avoid talking about my successes*), as this may modify the meaning of the items (in the example, *I avoid talking about my successes* would cease to be a marker of modesty and become a marker of extroversion). In any case, the predictive validity results of the TAPAS, whose adaptive test is based on this invariance, are positive (e.g., Trent et al., 2020).

Kreitchmann et al. (submitted for publication) found that an on-the-fly FC-CAT, under the assumption of invariance and an optimal selection procedure, shows a small improvement over an optimal bank-based FC-CAT (e.g., 0.01 in reliability coefficient), but large improvements in exposure control, since as the number of possible blocks increases it is more difficult for two individuals being evaluated to receive exactly the same blocks.

## DISCUSSION

Advances in technology and the development of psychometric models in the last two decades are making it possible to provide an answer to a classic problem: the measurement of personality in selection contexts in which social desirability may be high. In 2005, Salgado included forced-choice tests as a non-recommendable solution, partly because of the difficulty of analysis that this type of test involves. The evidence on their greater robustness to faking and greater predictive validity seems to tilt the balance towards a more positive view, provided that the problems of score ipsativity are resolved. Based on IRT, different models have been proposed (Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; Stark et al., 2005) that help to solve the problem of ipsativity. However, evaluation demands, block assembly, and block size can be of great importance. In general, test performance will be better the more dimensions are measured and the less correlated they are. In block assembly, not only is the matching of social desirability relevant, but also the contribution to reducing error variance. In this regard, whether to use heteropolar blocks to eliminate score ipsativity (Frick et al., 2021) or not (Morillo, 2018; Kreitchmann et al., 2021) is a matter of debate. For Bürkner et al. (2019), heteropolar blocks might be counterproductive in applied contexts. Our recommendation is the use of optimization algorithms to form homopolar blocks optimally. This may be challenging, as it requires the collection of information on the desirability of the items, as well as their prior calibration. However, we should not forget that: (a) the exploratory analysis of the structure may be more complex *a posteriori,* in two-dimensional blocks; and (b) non-optimal assembly will lead to problems relating to the ipsativity of scores. Finally, block size introduces added complexity in the creation of optimal blocks since, as block size increases, the number of possible blocks to choose from increases exponentially, making it less feasible to explore the universe of possibilities.

In short, although forced-choice tests have been around for a long time, their use has been reduced due to the limitations attributed to them. However, it is increasingly understood that this type of test is not a homogeneous category, and it is important to understand how the design of the test and the way it is scored influences its robustness to faking, the resolution of the ipsativity problem and, ultimately, its predictive validity. The most recent meta-analyses show that, in applied contexts, the use of quasi-ipsative forced-choice tests constitutes a promising strategy for obtaining greater predictive validity in the personality domain, with greater resistance to faking than other formats (Martinez & Salgado, 2021). Finally, the progress of new technologies and the development of new psychometric models are presented as two powerful allies that make it possible to construct adapted tests on the fly, optimizing test design and scoring.

## CONFLICT OF INTEREST

There is no conflict of interest.

## REFERENCES

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment, 25*(8), 959-977. https://doi.org/10.1177/1073191116667547

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

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460-502. https://doi.org/10.1177%2F0013164410375112

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology, 3*(4), 489-493. https://doi.org/10.1111/j.1754-9434.2010.01277.x

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior research methods, 44*(4), 1135-1147. https://doi.org/10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2018). Modelling forced-choice response formats. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 523-569). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch18

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827-854. https://doi.org/10.1177/0013164419832063

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347-1368. https://doi.org/10.1037/apl0000414

Chen, C., Wang, W., Chiu, M. M., & Ro, S. (2020). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement*, *57*(2), 343-369. https://doi.org/10.1111/jedm.12252

Cuadrado, D., Salgado, J. F., & Moscoso, S. (2021). Personality, intelligence, and counterproductive academic behaviors: A meta-analysis. *Journal of Personality and Social Psychology*, *120*(2), 504-537. https://doi.org/10.1037/pspp0000285

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army personnel selection and classification decisions.* Drasgow Consulting Group Urbana IL.

Fisher, P., Robie, C., Christiansen, N., Speer, A., & Schneider, L. (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus classical test theory scoring. *Personnel Assessment and Decisions*, *5*(1). https://doi.org/10.25035/pad.2019.01.003

Frick, S., Brown, A. & Eunike Wetzel (2021) Investigating the normativity of trait estimates from multidimensional forced-choice data, *Multivariate Behavioral Research*, https://doi.org/10.1080/00273171.2021.1938960

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*(1), 9-24. https://doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*(3), 167-184. https://doi.org/10.1037/h0029780

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598-612. https://doi.org/10.1177/0146621615585851

Hontangas, P. M., Leenen, I., & de la Torre, J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, *28*, 1, 76-82. https://doi.org/10.7334/psicothema2015.204

Judge, T. A., Rodell, 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(6), 875-925. https://doi.org/10.1037/a0033901

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology*, *10*, 2309. https://doi.org/10.3389/fpsyg.2019.02309

Kreitchmann, R. S., Abad, F. J., & Sorrel, M. A. (2021). A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01677-4

Kreitchmann, R. S., Sorrel, M. A., & Abad, F. J. (sent for publication). On bank assembly and block selection in multidimensional forced-choice adaptive assessments.

Lee, P., & Joo, S.-H. (2021). A new investigation of fake resistance of a multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions*, *7*(1). https://doi.org/10.25035/pad.2021.01.004

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389-414. https://doi.org/10.1177%2F0013164416646162

Makransky, G., Mortensen, E. L., & Glas, C. A. W. (2013). Improving personality facet scores with multidimensional computer adaptive testing: An illustration with the Neo Pi-R. *Assessment*, *20*(1), 3-13. https://doi.org/10.1177/1073191112437756

Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology*, *12*, 732241. https://doi.org/10.3389/fpsyg.2021.732241

Morillo, D. (2018). *Item response theory models for forced-choice questionnaires*. Doctoral dissertation, Universidad Autónoma de, Madrid.

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*(2), 75-83. https://doi.org/10.5093/jwop2019a11

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*(7), 500-516. https://doi.org/10.1177/0146621616662226

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*(2), 273-296. https://doi.org/10.1007/s11336-008-9097-5

Nieto, M. D., Abad, F. J., & Hernández-Camacho, A. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema*, *29.3*, 390-395. https://doi.org/10.7334/psicothema2016.391

Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five with bifactor computerized adaptive testing. *Psychological Assessment*, *30*(12), 1678-1690. https://doi.org/10.1037/pas0000631

Olea, J., Abad, F. J., & Barrada, J. R. (2010). Tests informatizados y otros nuevos tipos de tests [Computerized tests and other new types of test]. *Papeles del psicólogo*, *31*(1), 97-107.

Otero, I., Cuadrado, D., & Martínez, A. (2020). Convergent and predictive validity of the Big Five factors assessed with single stimulus and cuasi-ipsative questionnaires. *Journal of Work and Organizational Psychology*, *36*(3), 215-222. https://doi.org/10.5093/jwop2020a17

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences, 183*, 111114. https://doi.org/10.1016/j.paid.2021.111114

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338. https://doi.org/10.1037/a0014996

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347-364. https://doi.org/10.1177%2F107319110000700404

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387. https://doi.org/10.1037/a0026838

Salgado, J.F. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la psicología del trabajo y las organizaciones [Personality and social desirability in organizational settings: Practical implications for work and organizational psychology]. *Papeles del psicólogo, 92*, 115-128.

Salgado, J. F. (2016). A theoretical model of psychometric effects of faking on assessment procedures: Empirical findings and implications for personality at work: A Theoretical Model of faking psychometric effects. *International Journal of Selection and Assessment, 24*(3), 209-228. https://doi.org/10.1111/ijsa.12142

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*(4), 797-834. https://doi.org/10.1111/joop.12098

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*(1), 3-30. https://doi.org/10.1080/1359432X.2012.716198

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment, 27*(3), 572-584. https://doi.org/10.1177/1073191118762049

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*(3), 184-203. https://doi.org/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*(3), 153-164. https://doi.org/10.1037/mil0000044

Trent, J. D., Barron, L. G., Rose, M. R., & Carretta, T. R. (2020). Tailored Adaptive Personality Assessment System (TAPAS) as an indicator for counterproductive work behavior: Comparing validity in applicant, honest, and directed faking conditions. *Military Psychology, 32*(1), 51-59. https://doi.org/10.1080/08995605.2019.1652481

Tsutsui, S. (2006). Node histogram vs. edge histogram: A comparison of probabilistic model-building genetic algorithms in permutation domains. 2006 *IEEE International Conference on Evolutionary Computation*, 1939-1946. https://doi.org/10.1109/CEC.2006.1688a44

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197-210. https://doi.org/10.1177/00131649921969802