On Bank Assembly and Block Selection in Multidimensional Forced-Choice Adaptive Assessments

Rodrigo Schames Kreitchmann[1], Miguel A. Sorrel[1], Francisco J. Abad[1]

[1]Department of Social Psychology and Methodology, Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain

This is the author accepted author manuscript of the article published at Educational and Psychological Measurement, April 2023.

Kreitchmann, R. S., Sorrel, M. A., & Abad, F. J. (2023). On Bank Assembly and Block Selection in Multidimensional Forced-Choice Adaptive Assessments. Educational and Psychological Measurement, 83(2), 294-321. https://doi.org/10.1177/00131644221087986

Author Note

This project was partially supported by three grants from the Spanish Ministry of Economy, Industry and Competitiveness (projects PSI2015-65557-P, PSI2017-85022-P and FPI BES-2016-077814) and Cátedra de Modelos y Aplicaciones Psicométricas (Instituto de Ingeniería del Conocimiento and Autonomous University of Madrid).

Correspondence concerning this article should be addressed to Rodrigo S. Kreitchmann, Calle Iván Pavlov, 6, Ciudad Universitaria de Cantoblanco, 28049, Madrid, Spain. Telephone number: +34 914976138. E-mail: rodrigo.schames@uam.es.

Abstract

Multidimensional forced-choice (FC) questionnaires have been consistently found to reduce the effects of socially desirable responding and faking in non-cognitive assessments. Although FC has been considered problematic for providing ipsative scores under the classical test theory, IRT models enable the estimation of non-ipsative scores from FC responses. However, while some authors indicate that blocks composed of opposite-keyed items are necessary to retrieve normative scores, others suggest that these blocks may be less robust to faking, thus impairing the assessment validity. Accordingly, this article presents a simulation study to investigate whether it is possible to retrieve normative scores using only positively keyed items in pairwise FC computerized adaptive testing (CAT). Specifically, a simulation study addressed the effect of 1) different bank assembly (with a randomly assembled bank, an optimally assembled bank, and blocks assembled *on-the-fly* considering every possible pair of items), and 2) block selection rules (i.e., **T**, and Bayesian **D** and **A**-rules) over the estimate accuracy and ipsativity and overlap rates. Moreover, different questionnaire lengths (30 and 60) and trait structures (independent or positively correlated) were studied, and a non-adaptive questionnaire was included as baseline in each condition. In general, very good trait estimates were retrieved, despite using only positively keyed items. Although the best trait accuracy and lowest ipsativity were found using the Bayesian **A**-rule with questionnaires assembled *on-the-fly*, the **T**-rule under this method led to the worst results. This points out to the importance of considering both aspects when designing FC CAT.

Keywords: forced-choice format, ipsative data, multidimensional IRT, adaptive testing, item selection

On Bank Assembly and Block Selection in Multidimensional Forced-Choice Adaptive Assessments

Recent meta-analytic studies provide evidence for the predictive validity of non-cognitive domains, such as personality, motivation, and leadership over academic and job performances (e.g., Judge et al., 2013; Montano et al., 2017; Poropat, 2009; Richardson et al., 2012). As a result, there has been an increasing interest in the assessment of these attributes in educational and occupational fields. In this sense, forced-choice (FC) questionnaires have been proposed as a means of attenuating the effect of socially desirable responding (i.e., self-deception or faking) and acquiescence in the measurement of non-cognitive attributes (e.g., Cao & Drasgow, 2019; Cheung & Chan, 2002; Martínez & Salgado, 2021; Wetzel et al., 2021), thus providing with more valid measurements than with traditional rating scale items.

Differently from the traditional rating scale items (e.g., using Likert scales), the FC format consists of presenting the respondents with two or more stimuli (e.g., statements) in blocks. Respondents are then instructed to rank the statements within a block from *least like me* to *most like me*. Instructions may be either to rank all the statements (i.e., *full rank*), or to partially rank them (e.g., pick one, select only the most and least preferable, etc.; for an overview on different types of blocks see Brown & Maydeu-Olivares, 2018). In this sense, the simplest and most common FC format is with two statements, where respondents are instructed to pick the most preferable of each pair (e.g., Morillo et al., 2019).

A major drawback for the FC format, however, is that it may provide with ipsative scores, such that a person's score in each attribute depends on his own scores on other variables (e.g., scoring higher in one attribute implies scoring lower in the others), which prevents comparing respondents with each other. The interdependence between scale scores in FC response data is due to that endorsing the items measuring one attribute implies not endorsing the items measuring other attributes. One operational definition of pure ipsativity is

that the sum of scores across all traits is constant for all individuals. As a consequence of ipsativity, the validity of the scores is impaired. Specifically, pure ipsative scores hold a set of known unique psychometric properties (Clemans, 1966). For instance, due to the exact interdependence between the scales in a questionnaire, the sum of rows, or columns, in the trait covariance matrix must be zero, which also occurs with the trait intercorrelation matrix if trait variances are equal. Accordingly, in this case, the average trait intercorrelation between *D* ipsative scores from the same questionnaire is necessarily $-\frac{1}{D-1}$ (Hicks, 1970, p. 172), since the sum of the off-diagonal elements should be -1 by rows (or by columns). On the other hand, the sum of the covariances between a given external criterion and a set of ipsative variables will also be zero (Clemans, 1966, p. 28), which implies that, when ipsative variances are the same, the sum (and thus the average) of the correlations between the ipsative scores and a given external variable will also be zero (Clemans, 1966, p. 28). The aforementioned unique psychometric properties, such as the negative intercorrelation bias and a mean validity of zero, can be taken as score ipsativity indicators. The following is a demonstration of these properties:

Note that $\sum_{d=1}^{D} Cov(X_d, X_{d'}) = \sum_{d=1}^{D} \left(\frac{\sum_{i=1}^{N} x_{id} x_{id'}}{N} - \bar{x}_d \bar{x}_{d'} \right) =$

$$\sum_{i=1}^{N} \frac{(\sum_{d=1}^{D} X_{id}) X_{id'}}{N} - \bar{x}_{d'} \sum_{d=1}^{D} \bar{x}_d = K \sum_{i=1}^{N} \frac{x_{d'i}}{N} - \bar{x}_{d'} K = 0,$$

since $\sum_{d=1}^{D} X_{id} = \sum_{d=1}^{D} \bar{x}_d = K$, where *K* is a constant.

Note that $\sum_{d=1}^{D} Cov(X_d, y) = \sum_{d=1}^{D} \left(\frac{\sum_{i=1}^{N} (x_{id} y_i)}{N} - \bar{x}_d \bar{y} \right) =$

$$\sum_{i=1}^{N} \frac{(\sum_{d=1}^{D} X_{id})(y_i)}{N} - \bar{y} \sum_{d=1}^{D} \bar{x}_d = K \sum_{i=1}^{N} \frac{y_i}{N} - \bar{y} K = 0,$$

since $\sum_{d=1}^{D} X_{id} = \sum_{d=1}^{D} \bar{x}_d = K$, where *K* is a constant.

Modelling Forced-Choice Responses

Recently, it has become clear that the ipsativity of the scores can be attenuated by correctly modeling the process of comparative judgements (e.g., Meade, 2004). Nowadays, a variety of models under the item response theory (IRT) framework exist to outline the decision process involved in force-choice responses, thus allowing to obtain normative scores (e.g., Brown & Maydeu-Olivares, 2011; Bunji & Okada, 2020; McCloy et al., 2005; Morillo et al., 2016; Stark et al., 2005). In general, these models consist of two major components: (1) a measurement model for the relationships between stimuli and attributes, and (2) a decision model for the choice between the stimuli (Brown, 2016). A common denominator for most of these models, however, is the understanding that the endorsement of a given statement (i.e., item) in a block (i.e., set of items) results from the comparison of independent evaluations about each statement (Brown, 2016). On the other hand, these models mainly differ on their conception of the process underlying the individual item evaluations.

As an example for the comparative model component, the multi-unidimensional pairwise preference (MUPP) model proposed by Stark et al. (2005) outlines how the probability of endorsing an item in a pair, $P(y_{i,j})$ in Equation 1, results from an independent evaluation about the agreement with each item, $P(x_{i,j_1})$ and $P(x_{i,j_2})$.

$$P(y_{i,j}=1)=\frac{P(x_{i,j_1}=1)P(x_{i,j_2}=0)}{P(x_{i,j_1}=1)P(x_{i,j_2}=0)+P(x_{i,j_1}=0)P(x_{i,j_2}=1)}, \quad (1)$$

where $y_{i,j}$ represents the position of the selected item on the block (i.e., 1 or 2), and x_{i,j_1} and x_{i,j_2} denote the latent responses of subject *i* to items j_1 and j_2 in the j^{th} pair, respectively, being equal to 1 if respondent *i* endorses the item, and 0 otherwise.

Assuming the process outlined in Equation 1, a dominance pairwise preference model can be derived by specifying $P(x_{i,j_1})$ and $P(x_{i,j_2})$ through a two-parameter logistic (2PL) model defined in Equation 2:

$$P\left(x_{i,j_p} = 1 \middle| \theta_{i,d_{j_p}}\right) = \psi_{\text{logistic}}\left[a_{j_p}\left(\theta_{i,d_{j_p}} - b_{j_p}\right)\right], \tag{2}$$

where the probability of agreement with each statement p in the j^{th} pair depends on the discrimination and difficulty of the p^{th} item (a_{j_p} and b_{j_p}, respectively) and the i^{th} person score in the d^{th} latent trait ($\theta_{i,d_{j_p}}$) under the logistic link function (ψ_{logistic}). To ease the comprehension of upcoming definitions, Equation 2 can be redefined into the slope and intercept parametrization as

$$P\left(x_{i,j_p} = 1 \middle| \theta_{i,d_{j_p}}\right) = \psi_{\text{logistic}}\left(a_{j_p}\theta_{i,d_{j_p}} + c_{j_p}\right), \tag{3}$$

where $c_{j_p} = -a_{j_p}b_{j_p}$ is the j_p item's intercept.

By replacing the $P\left(x_{i,j_p}\right)$ terms in Equation 1 by the probability function of the 2PL (Equation 3), the MUPP-2PL (Morillo et al., 2016) simplifies to the logistic difference between parameters in $P\left(x_{i,j_1} = 1 \middle| \theta_{i,d_{j_1}}\right)$ and $P\left(x_{i,j_2} = 1 \middle| \theta_{i,d_{j_2}}\right)$:

$$P(y_{i,j} = 1|\boldsymbol{\theta}_i) = \psi_{\text{logistic}}\left[\left(\mathbf{a}'_{j_1}\boldsymbol{\theta}_i + c_{j_1}\right) - \left(\mathbf{a}'_{j_2}\boldsymbol{\theta}_i + c_{j_2}\right)\right] = \psi_{\text{logistic}}\left(\mathbf{s}'_j\boldsymbol{\theta}_i + c_j\right), \tag{4}$$

where $\boldsymbol{\theta}_i$ is a $D \times 1$ vector containing the true attribute levels of the i^{th} subject. Additionally, let $\mathbf{a}'_{j_p}$ be a $1 \times D$ vector of discrimination parameters of item j_p over the D attributes, being $a_{j_p,d} = 0$ if item j_p does not measure attribute d, the j^{th} block scale parameter vector $\mathbf{s}'_j$ is then the difference between discrimination vectors for the first and second items in the block (i.e., $\mathbf{s}'_j = \mathbf{a}'_{j_1} - \mathbf{a}'_{j_2}$). Similarly, parameter c_j denotes the block threshold parameter, computed as the difference between item intercepts, being $c_j = c_{j_1} - c_{j_2}$. The detailed simplification from Equation 1 to Equation 4 can be found in Kreitchmann et al. (2021).

As it can be noted in Equation 4, the MUPP-2PL model is equivalent to the multidimensional compensatory logistic model (MCLM; McKinley & Reckase, 1982) for the difference of item response parameters. In practical terms, if both items are keyed in the same direction, the highest the θ measured by j_2, the lowest is the probability of endorsing item 1, and vice-versa. Similarly, from a Thurstonian perspective of comparative judgment, the decision of endorsement of either statement relies on the difference between the utilities of the items (Thurstone, 1927) which monotonically increase with the attribute levels. In this sense, the MUPP-2PL and the Thurstonian IRT model (TIRT; Brown & Maydeu-Olivares, 2011) are different parametrizations of the same model (logistic and normal link functions, respectively). In this study, the MUPP-2PL will be used, although, given the great similarity between the MUPP-2PL and the TIRT models, the conclusions here may be largely generalizable to the TIRT model for pairs.

Challenges for the FC Format

Although the ipsativity of FC scores may be attenuated by properly modelling the response process, some questionnaire characteristics can still lead to remnant ipsativity, undermining the precision and validity of the estimated scores (e.g., Frick et al., 2021). For instance, under the dominance framework, the **S** matrix (a $J \times D$ matrix with the **s** vectors of all blocks as defined in Equation 4) should be of full rank for the model to be identified (Brown, 2016). In addition, other aspects of the questionnaire design have been found to affect the reliability of the scores. For instance, Brown and Maydeu-Olivares (2011; 2018) indicate that more reliable score estimates can be obtained by: 1) including blocks of items with different keyed directions, 2) increasing the number of traits being measured, 3) assessing traits with more negative average intercorrelations (for positive or mixed item keying), and 4) increasing the number of items per block. Additionally, in a comprehensive study, Frick et al. (2021) observed that questionnaires only including blocks of items keyed in

the same direction (i.e., homopolar blocks) were generally problematic, providing with ipsative scores and reliabilities below acceptable levels. In this sense, within the non-adaptive assessment framework, Kreitchmann et al. (2021) found that the ipsativity derived from using only homopolar blocks may be reduced by optimizing the block assembly. Specifically, these authors found that even when using homopolar pairs only, optimizing the questionnaires by minimizing the asymptotic variances of the trait score estimators could improve substantially the recovery of the scores and provide with acceptable score precision. The effect of block assembly, however, is still unknown in adaptive assessment applications.

Regarding the use of blocks with opposite-keyed items, there is currently an open debate (e.g., Bürkner et al., 2019; Lee & Joo, 2021; Ng et al., 2021) on whether, despite improving ipsativity, it may reduce the robustness of FC blocks against faking and social desirability, again compromising validity. On the one hand, Bürkner et al. (2019) postulate that if traits are oriented in the same direction as social desirability, positively keyed items will most likely be more socially desirable, whereas negatively keyed items will be undesirable. It is necessary to emphasize that the item keying referred here is relative to the direction of the social desirability, and not the original direction of the measured domain. Therefore, unequally keyed blocks are likely to have a clear socially desirable response and thus fail to control social desirability and faking. Additionally, in realistic scenarios, if respondents are able to identify and select the most socially desirable option in a block, that block will be uninformative for person parameter estimation (Wang et al., 2017) and may not improve the accuracy of trait estimates as initially expected. On the other hand, Wetzel et al. (2021) found that FC questionnaires containing blocks of opposite-keyed items were still more robust to faking than rating scale items. In this sense, it appears to be assumable that, when items truly have the same social desirability for respondents, the inclusion of opposite-keyed items may solve ipsativity issues without compromising the validity of the assessment.

From a conservative standing point, however, it can be postulated that while the reliability and ipsativity of questionnaires using only equally-keyed blocks may be improved using longer or properly optimized questionnaires (e.g., Kreitchmann et al., 2021), the possible validity impairment caused by faking and social desirability may still be difficult to predict and prevent. Accordingly, this study aims to investigate whether, under realistic conditions, it is possible to retrieve reliable and valid scores with adaptive questionnaires using only blocks of items keyed in the same direction.

Adaptive Forced-Choice Assessment

Despite the challenges stated above, the use of computerized adaptive testing (CAT) is known to offer a substantial increment in trait estimate precision, which may likely provide with reliable normative scores. Accordingly, several CAT applications exist for measuring non-cognitive traits with the FC format, for instance the Navy Computer Adaptive Personality Scales (NCAPS; Houston et al., 2006), the Global Personality Inventory - Adaptive (GPI-A; CEB, 2010), or the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow et al., 2012; Stark et al., 2014).

Block Bank Assembly

In the FC framework, the assembly of block banks has its peculiarities. On the one hand, banks may be composed of independent blocks as in traditional CAT, by assigning each item to a single block, which allows for the calibration of the block bank with FC responses before the FC CAT implementation. Under this method, the maximum bank size will be M/V, being M the size of the item pool and V the number of items per block. Alternatively, blocks may be assembled *on-the-fly*, in the sense that, given the item parameters obtained from the calibration with single-stimulus responses, the block parameters can be anticipated as in Equation 4. Therefore, the suitability of each possible

combination of items can be calculated for each respondent and the best blocks are administered. TAPAS and NCAPS, for instance, are examples of *on-the-fly* applications. Compared with traditional banks, *on-the-fly* implementations have a much larger block search space, including thousands of possible blocks. For multidimensional FC CAT *on-the-fly*, the size of the search space (*B*) is defined in Equation 5.

$$B = \binom{M}{V} - \sum_{d=1}^{D} \binom{M_d}{V}, \quad (5)$$

where $\binom{M}{V}$ denotes the total combinations of the *M* items in the pool in blocks with size *V*, and $\binom{M_d}{V}$ indicates the number of unidimensional blocks formed with the items measuring each d^{th} dimension. For instance, for the 240-item bank measuring 5 dimensions, with 48 items per dimension, the total number of possible pairwise combinations is $\binom{240}{2}$ = 28,680, which reduces to *B* = 23,040 after excluding the $\sum_{d=1}^{5} \binom{48}{2}$ = 5,640 unidimensional pairs. Although unidimensional pairs may be included, it was the authors' decision not to include them in this study. In this sense, as can be inferred from Equation 4, unidimensional homopolar blocks may be little informative since the block discrimination over dimension *d* will be the subtraction of the discriminations of the two items regarding this dimension. In order to be informative, unidimensional blocks must be composed of items with very divergent discriminations in that dimension, which may make it easier to identify the most desirable response (Cao & Drasgow, 2019, p. 1349).

Each bank assembly procedure has its pros and cons. For instance, due to its reduced size and the conditional independence between blocks (if items are not repeated in multiple blocks), traditional (i.e., fixed) block banks allow for the calibration with forced-choice responses, and thus enable the inspection of the psychometric properties of the blocks before the FC CAT implementation. On the other hand, CATs assembled *on-the-fly* provide with a

much larger search space and many more suitable blocks for each $\boldsymbol{\theta}$ profile, improving the measurement for all subjects. Nevertheless, an actual calibration with FC response data is not feasible for blocks assembled *on-the-fly*, which may have unpredictable effects on the validity of the assessment. For instance, while there is evidence of the invariance from rating scales to FC parameters when respondents are expected to respond honestly (Lin & Brown, 2017; Morillo et al., 2019), the invariance does not necessarily hold when respondents try to fake (Lee & Joo, 2021). In this sense, if a block has a more desirable option in situations in which faking is expected, assuming the parameters from the single-stimulus item calibration may provide with artificially inflated scores in the dimension measured by the more desirable items. This, in turn, may also compromise the fairness of the assessment, as these more fakeable blocks may not be presented systematically to all respondents. Finally, whereas FC CATs formed *on-the-fly* account for the full combinatorics of multidimensional blocks, the choice on how to combine the items to assemble the reduced fixed banks may affect the performance of the assessment. For instance, as it is known from previous literature (e.g., Brown, 2016) and will be further detailed, the combination of item scale parameters in the blocks may determine the normativity of the estimated scores. Conversely, given the breadth of the search space in FC CATs formed *on-the-fly*, the performance of the FC CAT *on-the-fly* may be greatly influenced by the block selection rules, as there may be a margin for suboptimal selection rules to present respondents with only suboptimal blocks.

There is currently a lack of studies about the effect of the different FC CAT bank assembly procedures (i.e., using a fixed block bank *versus* assembling blocks *on-the-fly*) over the precision of the score estimates. In addition, as it will be further detailed, different selection rules may present respondents with blocks with different scale parameters combinations, thus possibly affecting the normativity/ipsativity of the scores (Brown & Maydeu-Olivares, 2018; Morillo, 2018, pp. 63-100). Therefore, it is also crucial to investigate how the selection rules

may affect the ipsativity (consequently, the predictive validity) of the FC scores in each bank assembly method. Finally, item and block exposure with the different selection rules and FC CAT assembly methods should be considered as they may affect test security, since overexposed items or blocks may allow respondents to know the "ideal" response beforehand (Chang, 2004). Specifically, overly exposed items or blocks may facilitate test security breaches, as they are more likely to be remembered by the respondents. If the blocks are known beforehand, future respondents may receive specific training on how to respond to achieve the desired score.

Block Selection Rules

In general terms, the Fisher information function is the starting point for most CAT selection rules, as it quantifies the information about the unknown $\boldsymbol{\theta}$ in the observations (Mulder & van der Linden, 2009). In unidimensional CAT, the item selection can be very straightforward, since choosing items with the maximum Fisher information for the respondent's latest θ estimate is equivalent in practice to minimizing the asymptotic standard error of that $\hat{\theta}$. In multidimensional FC contexts, however, the Fisher information function becomes a *D*-dimensional square matrix and the different decomposition methods may lead to different outcomes. For instance, in a two-dimensional questionnaire (i.e., $D = 2$), the Fisher information function of block *j* under the MUPP-2PL model (as in Morillo et al., 2016) is given by:

$$\mathbf{I}_j(\boldsymbol{\theta}) = \begin{bmatrix} s_{j,1}^2 P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) & s_{j,1} s_{j,2} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) \\ s_{j,1} s_{j,2} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) & s_{j,2}^2 P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) \end{bmatrix}. \quad (6)$$

where $P_j(\boldsymbol{\theta})$ is the probability of endorsement of the first item in the j^{th} block, as in Equation 4, and $Q_j(\boldsymbol{\theta}) = 1 - P_j(\boldsymbol{\theta})$. Under the assumption of conditional independence between blocks (i.e., without repeating items in block), the test information matrix becomes:

$$\mathbf{I}(\boldsymbol{\theta}) = \sum_{j=1}^{J} I_j(\boldsymbol{\theta}). \tag{7}$$

The asymptotic variance-covariance matrix for maximum-likelihood trait estimators can be defined as (van der Linden, 2006):

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}\right) = \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \frac{\sum_j s_{j,1}^2 P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})}{|\mathbf{I}(\boldsymbol{\theta})|} & \frac{\sum_j s_{j,1} s_{j,2} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})}{|\mathbf{I}(\boldsymbol{\theta})|} \\ \frac{\sum_j s_{j,1} s_{j,2} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})}{|\mathbf{I}(\boldsymbol{\theta})|} & \frac{\sum_j s_{j,2}^2 P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})}{|\mathbf{I}(\boldsymbol{\theta})|} \end{bmatrix}, \tag{8}$$

where

$$|\mathbf{I}(\boldsymbol{\theta})| = \left[\sum_j s_{j,1}^2 P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})\right]\left[\sum_j s_{j,2}^2 P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})\right] - \left[\sum_j s_{j,1} s_{j,2} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})\right]^2. \tag{9}$$

Several information-based selection rules exist for multidimensional CAT, representing different decomposition methods of $\mathbf{I}(\boldsymbol{\theta})$ or $\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}\right)$ (e.g., Luecht, 1996; van der Linden, 1999). For instance, blocks may be selected to maximize the trace of the Fisher information matrix (**T**-rule; e.g., Joo et al., 2020). However, as it can be observed in Equations 6 to 9 on a two-dimensional MUPP-2PL model, although the **T**-rule being proportional to the sum of $s_{j,1}^2$ and $s_{j,2}^2$, the asymptotic error variances are also related with the determinant of the information matrix, which in turn will be higher for lower $s_{j,1}s_{j,2}$ products. Consequently, selecting blocks with the **T**-rule does not necessarily reduce the error of $\widehat{\boldsymbol{\theta}}$. Alternatively, the **D**-rule consists of selecting blocks that iteratively maximize the determinant of the questionnaire information matrix at time *j*. Using the latest $\widehat{\boldsymbol{\theta}}$, the questionnaire information obtained by administering a given block *l* at time *j* can be anticipated by adding up the information provided by all previously applied blocks ($\mathbf{I}_{j-1}$) plus the information of block *l* at time *j*, thus the **D**-rule can be defined as Equation 10:

$$\text{maximize } \left|\mathbf{I}_{j-1}(\widehat{\boldsymbol{\theta}}^{j-1}) + \mathbf{I}_{j_l}(\widehat{\boldsymbol{\theta}}^{j-1})\right|. \quad (10)$$

Similarly, this approach can be extended to iteratively minimize the sum of the asymptotic variances of $\widehat{\boldsymbol{\theta}}$ for each questionnaire with length j (i.e., minimizing the trace of $\mathbf{I}_j^{-1}$; **A**-rule):

$$\text{minimize } \text{tr}\left[\mathbf{I}_{j-1}(\widehat{\boldsymbol{\theta}}^{j-1}) + \mathbf{I}_{j_l}(\widehat{\boldsymbol{\theta}}^{j-1})\right]^{-1} \quad (11)$$

As pointed out by Segall (1996), the use of the posterior information matrix instead of Fisher information matrix improves the efficiency of Bayesian $\boldsymbol{\theta}$ estimators by accounting for prior information of $\boldsymbol{\theta}$. Thus, Bayesian **D** and **A**-rules (noted here as $\mathbf{D}^*$ and $\mathbf{A}^*$) can be defined by using the posterior information function ($\mathbf{I}_j^*$) by adding the inverse of the trait variance-covariance matrix ($\boldsymbol{\Phi}$) to the questionnaire information matrix (see Equation 12). Please note that Bayesian and non-Bayesian **T**-rule are equivalent, since adding constant $\boldsymbol{\Phi}^{-1}$ to every block information matrix will lead to the same mode of $\mathbf{I}_{j_l}(\widehat{\boldsymbol{\theta}}^{j-1})$, thus selecting the same block *l*. A comprehensive comparison between **T**, **D** and **A**-rules can be found in Mulder and van der Linden (2009).

$$\mathbf{I}_{j_l}^* = \mathbf{I}_{j-1}(\widehat{\boldsymbol{\theta}}^{j-1}) + \mathbf{I}_{j_l}(\widehat{\boldsymbol{\theta}}^{j-1}) + \boldsymbol{\Phi}^{-1} \quad (12)$$

Goals of the Present Study

The main goal of this study is to investigate whether adaptive assessments can facilitate the recovery of normative scores with questionnaires composed only with homopolar item pairs, as well as to identify the necessary conditions for its occurrence. Specifically, this study addresses, through simulation, the effect of 1) bank assembly procedures (a fixed randomly assembled bank, a fixed optimally assembled bank, and blocks assembled *on-the-fly*), and 2) block selection rules (i.e., **T**, and Bayesian $\mathbf{D}^*$ and $\mathbf{A}^*$-rules) over the accuracy and ipsativity of trait scores, as well as overlap rates. Additionally, different questionnaire lengths and trait structures (independent or positively correlated) are studied, and a non-adaptive questionnaire is included as baseline in each condition.

Simulation Study

Method

Item Pool Generation

Five-dimensional 240-item pools with 48 items per dimension were simulated to emulate a real FC personality CAT under the five-factor model (Costa & McCrae, 1992). Two FC CAT lengths were defined: J = 30 and 60, with 6 and 12 items per dimension, respectively. For each item pool, the item discrimination parameters (a_{j_p}) were sampled from a U(0.5, 2.5) distribution, and item difficulty parameters (b_{j_p}) from a U(-2.0, 2.0) distribution (c_{j_p} values were later calculated as $c_{j_p} = -a_{j_p} b_{j_p}$). Twenty replications were conducted for each condition. As it is often done in CAT studies (e.g., Mulder & van der Linden, 2009), the item bank is assumed to be precisely calibrated, thus the true model parameters were taken as known to compute the trait scores. This allows to compute the upper-limit performance of the adaptive assessment and what is expected to obtain in practical settings provided the item parameters are accurately estimated (e.g., Sorrel et al., 2021). The *R* codes used for data generation and the simulation study are available from the corresponding author upon request.

Block Bank Assembly

As previously mentioned, three bank assembly procedures were considered: 1) randomly assembling a fixed block bank, 2) optimally forming a fixed block bank, and 3) forming blocks *on-the-fly*. For the first, a fixed 120-block pool was generated by randomly pairing all the items from the 240-item pool without repetition, with every block measuring two dimensions. For the optimal fixed bank, items were paired using a genetic algorithm to maximize the average posterior marginal reliability of $\boldsymbol{\theta}$ (Kreitchmann et al., 2021), also constrained to two-dimensional blocks and no item repetition. Differently, for FC CAT *on-*

the-fly, a pairwise FC block bank was defined with all possible two-dimensional item pairs, with length *B* defined by Equation 12.

$$B = \binom{M}{2} - \sum_{d=1}^{D} \binom{M_d}{2} \quad (12)$$

where $\binom{M}{2}$ denotes the total of pairwise combinations of the *M* items in the pool, and $\binom{M_d}{2}$ indicates the number of unidimensional pairs formed with the items addressing each d^{th} dimension. For the 240-item bank measuring 5 dimensions with 48 items per dimension, the number of possible (both one and two-dimensional) pairwise combinations was $\binom{240}{2}$ = 28,680, which reduced to *B* = 23,040 after excluding the $\sum_{d=1}^{5}\binom{48}{2}$ = 5,640 unidimensional pairs.

Respondent Data Generation

As in Brown and Maydeu-Olivares (2011), the true latent trait correlation matrix ($\boldsymbol{\Phi}$) was set as either a five-dimensional identity matrix ($\boldsymbol{I}_5$) or as the one observed for the revised NEO personality inventory (NEO PI-R; Costa & McCrae, 1992) with empirical data (see Table 1). To analyze the recovery of the trait estimates in each condition and replication, true $\boldsymbol{\theta}$ were drawn from $MVN(\mathbf{0}, \boldsymbol{\Phi})$ for 1,000 simulees. Forced-choice response data were then generated given the probabilities under the MUPP-2PL model using the true parameters of the selected blocks for each FC CAT or non-adaptive questionnaire analyzed. Aiming to investigate effect of remnant ipsativity over the criterion validity of the scores, an external variable ξ with standard normal distribution was simulated with correlation of 0.3 with $\boldsymbol{\theta}$.

Please insert Table 1 here

Block Selection Rules

The adaptive algorithm was defined to present each respondent with a tailored assessment with length *J* under each selection rule and FC CAT assembly method. An initial set of blocks addressing each of the dimensions measured is required to obtain the first trait estimates (e.g., Mulder & van der Linden, 2009). For greater comparability of the block selection rules performance and aiming not to inflate the overlap rates due to the selection of the same blocks in the beginning of the FC CAT, the first three blocks administered to each examinee were chosen at random (e.g., Kaplan et al., 2015), constrained to having at least one block measuring each dimension (e.g., 1-2, 3-4, 4-5), and these three initial blocks were fixed across the different selection rules. A schematic description of the adaptive algorithm is presented in Figure 1.

Please insert Figure 1 here

The non-adaptive FC questionnaires used as baseline comparison were generated by assembling the 240 simulated items into *J* pairs using a genetic algorithm that maximizes the average posterior marginal reliability of $\boldsymbol{\theta}$ (Kreitchmann et al., 2021). As described by Kreitchmann et al. (2021), the optimized criterion is inversely proportional to a weighted $\mathbf{A}^*$-rule, minimizing $\sum_q tr[\mathbf{I}(\boldsymbol{\theta}_q)^{-1}]\, w(\boldsymbol{\theta}_q)$ across a set of *Q* $\boldsymbol{\theta}$ vectors, being $w(\boldsymbol{\theta}_q)$ the normalized density of vector $\boldsymbol{\theta}_q$, i.e., $\sum_q w(\boldsymbol{\theta}_q) = 1$. Under all conditions, the number of blocks measuring each pair of dimensions, among the $\binom{5}{2} = 10$ dimension pairs, was constrained to be equal for all dimension pairs (3 and 6 for *J* = 30 and 60, respectively).

Analyses

Trait Score Accuracy. For each simulated data set, the trait scores were estimated using *Maximum-a-Posteriori* method with Quasi-Monte Carlo quadrature in *mirt* package (Chalmers, 2012) using the true model parameters. Two indices were calculated to evaluate de recovery of trait scores: 1) the true reliability, calculated through the squared correlation between true and estimated θ ($r^2_{\theta\hat{\theta}}$), and 2) the root-mean-square error between estimated and true θ ($\text{RMSE}_{\hat{\theta}}$; Equation 13). Both $\text{RMSE}_{\hat{\theta}}$ and $r^2_{\theta\hat{\theta}}$ were computed for each dimension separately and then averaged across the five traits. In addition, the shape of the average conditional standard errors of θ under each assembly method was graphically analyzed for θ values between -2 and 2.

$$\text{RMSE}_{\hat{\theta}} = \sqrt{\sum_{n=1}^{N} \frac{\left(\hat{\theta}_n - \theta_n\right)^2}{N}}, \tag{13}$$

where N denotes the simulated sample size.

Trait Score Ipsativity. As indicated in the introduction section, ipsative scores have a set of unique psychometric properties that affect the the validity of the assessments. In this sense, aiming to approximate the ipsativity of the scores, two indicators were calculated to quantify the degree to which the validity of the assessment are impaired: 1) the trait intercorrelation bias ($\text{Bias}_{\hat{\mathbf{\Phi}}}$; Equation 14), and 2) the average correlation between $\hat{\boldsymbol{\theta}}$ and the simulated ξ criterion ($\bar{r}_{\xi\hat{\boldsymbol{\theta}}}$), disattenuated dividing by the true reliabilities $r^2_{\theta\hat{\theta}}$.

$$\text{Bias}_{\hat{\mathbf{\Phi}}} = \hat{\mathbf{\Phi}} - \mathbf{\Phi} \tag{14}$$

The $\text{Bias}_{\hat{\mathbf{\Phi}}}$ and $\bar{r}_{\xi\hat{\boldsymbol{\theta}}}$ were Fisher Z-transformed prior to calculating correlation differences or means across the five traits, and later backtransformed to the correlation metric (e.g., Corey et al., 1998).

Item and Block Exposure. As previously mentioned, knowing how the selection rules may affect item and block exposure is crucial in the search for valid educational and personnel selection assessment, as overexposed items or blocks may compromise test security, thus allowing respondents to know the "ideal" response beforehand (Chang, 2004). In this sense, item ($\bar{T}_M$) overlap rates (Equation 15) were used to quantify the average proportion of items that are shared by two random respondents (Chen et al., 2003). Similarly, the calculation of block overlap rates ($\bar{T}_B$; Equation 16) was adapted from Equation 15.

$$\bar{T}_M = \frac{M}{K \cdot J} \cdot S^2_{r_M} + \frac{K \cdot J}{M} \tag{15}$$

$$\bar{T}_B = \frac{B}{J} \cdot S^2_{r_B} + \frac{J}{B} \tag{16}$$

where M and B are the item and block bank sizes (i.e., M = 240, and B = 120 or 23,040), K is the block size (i.e., K = 2), and $S^2_{r_M}$ and $S^2_{r_B}$ are the variances of the item and block exposure rates across all items/blocks. In addition, as can be inferred from Equations 15 and 16, high overlap rates are also indicative of extreme usage pattern (i.e., with high exposure variances, thus with few overexposed items).

Effect Sizes. Finally, the effects of the manipulated factors over each of the indicators were synthesized through multiple analyses of variance (ANOVAs), one for each assessment procedure (i.e., non-adaptive test, FC CAT with a fixed random bank, FC CAT with a fixed optimal bank, and FC CAT *on-the-fly*) and indicator. For the adaptive assessment applications, the selection rule was outlined as a within-group factor in mixed-effects ANOVAs. Partial η^2 and generalized η^2 (Olejnik & Algina, 2003) effect sizes were used to quantify the relevance of these effects in fixed-effects and mixed-effects ANOVAs, respectively. All analyses were conducted using *R* software (R Core Team, 2020) and

ANOVAs were performed with the Type III sum of squares using the *afex* package (Singmann et al., 2020).

Results

Trait Score Accuracy

The values of the trait estimate accuracy indicators are listed in Table 2. In accordance with the previous literature (e.g., Brown & Maydeu-Olivares, 2011), the number of items per trait (and, consequently, the test length) and the true trait intercorrelations are expected to affect the accuracy of the scores. Accordingly, across all assessment methods, the score precision was greatly affected by the questionnaire length (η^2 from 0.93 to 0.98, see Table 3), with a minimum of $r^2_{\theta\hat{\theta}} = 0.63$ using the **T**-rule in FC CAT *on-the-fly* with $J = 30$ and $\mathbf{\Phi}$ = NEO PI-R, and a maximum of $r^2_{\theta\hat{\theta}} = 0.91$ with the $\mathbf{A}^*$-rule in FC CAT *on-the-fly* with $J = 60$. The true trait intercorrelations, on the other hand, had an effect over the accuracy of the scores only under adaptive measurements, and depended further on the pairing procedure (random or optimal) used to form the fixed banks, and on the selection rules under *on-the-fly* applications. Specifically, the reduced accuracy caused by the average positive true trait intercorrelations (i.e., $\mathbf{\Phi}$ = NEO PI-R) was negligible when fixed questionnaires or fixed block banks were optimized to maximize the average marginal posterior reliabilities. Additionally, regarding CATs assembled *on-the-fly*, the recovery of the scores in questionnaires using the $\mathbf{A}^*$-rule was less affected by $\mathbf{\Phi}$ than the other rules.

Similarly, the selection rule alone had an important impact over the accuracy of the scores in adaptive assessments. In this sense, the $\mathbf{A}^*$-rule consistently provided with the most precise $\boldsymbol{\theta}$ estimates, whereas the **T**-rule provided with the most inaccurate. As anticipated in the Introduction section, such effect was stronger in *on-the-fly* assessments (η^2 from 0.98 to 0.99) as the size of the combinatorial space gave a greater margin for the selection of many

more suboptimal blocks under suboptimal rules (i.e., **T** and **D***-rule). Accordingly, as it will be described further, the profile of scale parameter combinations (e.g., both a_{j_p} high, or one a_{j_p} and the other a_{j_p} low) of the selected blocks differed under each rule. In summary, using **A***-rule provided with the most accurate trait scores regardless of the adaptive bank assembly method. Under this rule, questionnaires assembled *on-the-fly* offered very high score precision, with $r^2_{\theta\hat{\theta}}$ ranging between 0.84 and 0.91.

Please insert Table 2 here

Please insert Table 3 here

Figure 2 illustrates the shape of the distribution of average conditional standard errors of $\boldsymbol{\theta}$ for each assessment method and $\boldsymbol{\Phi}$ with **A***-rule and J = 60. Non-adaptive assessments and CATs with optimally assembled banks offered higher average standard errors for extreme $\boldsymbol{\theta}$, which may be due to the optimization criterion used in the genetic algorithm (maximize the marginal posterior reliabilities). As exposed in Kreitchmann et al. (2021), the criterion used considers the density of $\boldsymbol{\theta}$ for the optimization, which leads to lower standard errors for the most populated $\boldsymbol{\theta}$ values. Although this may be manipulated while optimizing a fixed test or a block bank (e.g., van der Linden, 1996), o*n-the-fly* implementations showed better results without the necessity of prespecifying the desired shape for the standard errors. Finally, the shapes of the average conditional standard error distribution were similar in the remaining conditions, so they were omitted from the figure.

Please insert Figure 2 here

Trait Score Ipsativity

The average results for the trait intercorrelation bias and the estimated criterion validity of the scores under the different assessment conditions are presented in Table 4. In general, Tables 2 and 4 present a very similar pattern, thus the table with effect sizes is omitted for the ipsativity indicators. The best score ipsativity results ($\text{Bias}_{\widehat{\mathbf{\Phi}}}$ closer to 0 and $r_{\xi\widehat{\theta}}$ closer to 0.3) were obtained with the $\mathbf{A}^{*}$-rule in FC CAT *on-the-fly*. Similar to the accuracy results, the pairing procedure (random or optimal) had an important effect over the ipsativity indicators in CATs with fixed banks (η^2 = 0.53 for $r_{\xi\widehat{\theta}}$ and η^2 = 0.77 for $\text{Bias}_{\widehat{\mathbf{\Phi}}}$), while the selection rule had a major effect in applications *on-the-fly*. Additionally, the **T**-rule also offered the worst ipsativity results. In this sense, although reliability was acceptable in some conditions ($r^2_{\theta\widehat{\theta}} = 0.74$ to 0.78 with **Φ** = Identity), FC CATs *on-the-fly* with the **T**-rule provided with close to fully ipsative scores (i.e., $\text{Bias}_{\widehat{\mathbf{\Phi}}}$ close to $-1/(D-1) = -0.25$ and average criterion validity close to 0).

Please insert Table 4 here

Item and Block Exposure

The item and block overlap rates are listed in Table 5. As it can be inferred, overlap rates for the non-adaptive assessment were always 1.0 and were omitted from the table, since the same questionnaires were used to score all the respondents. Additionally, item and block overlap rates are the same in fixed bank FC CAT implementations, as each item is presented only in one block. Also as expected, the overlap rates increased with questionnaire length under adaptive assessments, as the ratio between FC CAT length and pool size increased (see Equations 15 and 16). Regarding the difference between adaptive assessment methods, item overlap was greater for FC CAT *on-the-fly*, indicating that a small proportion of the items

was administered with a greater frequency. Nonetheless, block overlap was substantially lower for *on-the-fly* implementations, indicating that although some items were overexposed, they were paired differently for each respondent. Additionally, a difference in overlap rates for random and optimal fixed banks was observed, for which optimizing the bank assembly also offered an advantage in overlap, regardless of the selection rule used.

Please insert Table 5 here

On the one hand, as it can be inferred from Equations 6 to 9, the better overall performance of the **A***-rule in FC CATs *on-the-fly* may be associated with maximizing the scale parameters for each dimension separately throughout the FC CAT while pairing items with different scale parameters in each block (e.g., for the two-dimensional case, maximizing $s_{j,1}$ and $s_{j,2}$ in different blocks, while minimizing the product $s_{j,1}s_{j,2}$). On the other hand, the worse results for **T**-rule may relate with the fact that it maximizes all the scale parameters in each block, thus providing with higher estimator variances. To illustrate it, Figure 3 presents the distribution of item scale parameters for the most exposed blocks for FC CATs *on-the-fly* under each selection rule in one replication for the 60-block condition. Accordingly, the high item overlaps in Table 5 reflect the fact that an important part of the item bank is less frequently administered under either of the selection rules. For the **A***-rule, for instance, item pool design including both low and high scale parameters should help to reduce the item overlap rate.

Please insert Figure 3 here

Discussion

Due to the risk of ipsativity by using only homopolar blocks in FC questionnaires, the implementation of adaptive assessments may be crucial to retrieve reliable and valid FC scores. In addition, different bank assembly procedures and information-based block

selection rules may derive into different assessment reliabilities and validities. Accordingly, this study aimed to investigate the effect of different adaptive assessment methods (i.e., with a fixed randomly formed FC bank, with a fixed optimally assembled FC bank, or with blocks assembled *on-the-fly*), and selection rules (i.e., **T**, **D*** and **A***-rules) on the accuracy and validity of the scores they provide, as well as their impact on test security.

Main Findings

As a general summary, this study provided with empirical evidence that, with the proper optimization criterion (**A***-rule), it is possible to retrieve reliable and valid scores using only homopolar FC blocks. Consistent with CAT literature (e.g., Mulder & van der Linden, 2009), the selection rules had a substantial effect over trait recovery under adaptive assessments, especially in FC CAT *on-the-fly*, being **T**-rule and **A***-rule the ones offering the worst and the best score precision, respectively. In addition, according to previous studies (e.g., Brown & Maydeu-Olivares, 2011; Brown & Maydeu-Olivares, 2018; Frick et al., 2021) even better results should be expected from questionnaires in more favorable conditions (i.e., measuring more traits, or including more items per block).

Comparing with the other assessment methods, using FC CAT *on-the-fly* offered a substantial improvement in reliability, ipsativity and block overlap rates, also achieving low average standard errors of $\widehat{\boldsymbol{\theta}}$ throughout the $\boldsymbol{\theta}$ continuum. On the other hand, item overlap rates were higher with this FC CAT format. It can be hypothesized, however, that the "ideal" response in the pairwise FC format depends on both items in a pair. Therefore, a high item overlap may not compromise test security, as knowing individual items beforehand may not facilitate faking block responses. Accordingly, low block overlap rates should be pursued to enhance FC test security. Additionally, the main limitation for FC CAT *on-the-fly* is that the

parametric invariance from graded response (Equation 3) to FC blocks (Equation 4) must be assumed. Despite the existing evidence supporting the invariance assumption (Lin & Brown, 2017; Morillo et al., 2019), it is an important aspect to take into consideration. If there is any suspicion that this assumption may not hold, optimizing the blocks using the single-stimulus item parameters (as in Kreitchmann et al., 2021) and re-calibrating the block parameters with FC responses would allow to exclude any non-invariant block. Using non-optimized (i.e., randomly assembled) fixed banks, on the other hand, would not be recommended, as it provided with the worse ipsativity.

It should be noted that in general, the results found here were not entirely consistent with most previous simulation studies. Frick et al. (2021), for instance, found that questionnaires composed with all positively keyed items provided with biased trait estimates and did not achieve acceptable overall reliability and ipsativity. Low reliabilities were also found by Bürkner et al. (2019) when measuring up to five dimensions with only equally keyed items, although higher reliabilities were found when measuring thirty dimensions. Additionally, differently from the expected (e.g., Brown & Maydeu-Olivares, 2011; Frick et al., 2021), having positive average true trait intercorrelations (**Φ**) only affected the reliability and ipsativity of the scores when questionnaires were not properly optimized (i.e., with the randomly assembled fixed bank or FC CAT *on-the-fly* with the **T**-rule). These inconsistencies may indicate that the FC questionnaires with all positively keyed items used in previous studies differed from this study in some respects. Possible explanations for the inconsistencies may be that, in previous studies, the item combinations were not optimized (i.e., blocks were not made to minimize the standard errors of measurement), or the scale parameter/factor loading distributions were narrower. For instance, while Frick et al. (2021) simulated standardized loadings ranging from 0.65 to 0.95, the scale parameters generated in the present study ranged from 0.28 to 0.86 in factor analysis metric. Similarly, Schulte et al.

(2021) indicated that with lower and normally distributed factor loadings, the true reliabilities provided with the TIRT model were generally unacceptable. As indicated in Equations 8 and 9, the determinant of the Fisher information matrix, and consequently the asymptotic estimator variances, are greatly affected by the product of the item scale parameters. In other words, if item scale parameters are too similar, the determinant will be close to 0, and estimator variances will be too high. In this sense, the empirical data appears to support the fact that the scale parameters of personality items often have high variability. For instance, using the Big Five Triplets (Wetzel & Frick, 2020), Frick et al. (2021) found standardized loadings for the positively keyed items (with neuroticism reversed to emotional stability) ranging from 0.07 to 0.90 (approximately from 0.12 to 3.51 in a_{j_p} logistic IRT parametrization), with an average of 0.60 (average a_{j_p} of 1.26). As suggested by a reviewer, a follow up simulation was conducted to investigate the effect of the bank assembly methods and block selection rules with normally distributed item discrimination. Specifically, the parameter distribution proposed by Schulte et al. (2021) was replicated, with factor loadings drawn from a truncated $N(0.50, 0.16)$ within the limits of 0.1 and 0.9. The results can be found at https://osf.io/tyhfk/. As a summary, the true reliabilities were lower with the normal discrimination parameter distribution, although they were still generally acceptable [comparing Tables 2 (uniformly distributed item parameters) and S2 (normally distributed item parameters) the decrease in average reliability was small, 0.044 with $SD = 0.017$]. In comparison with the results with uniform item discriminations, the effect of optimizing the questionnaires through bank assembly or block selection rules had a lower impact on the normativity of the questionnaires. The main conclusions of this article, however, remained the same: 1) the forced-choice questionnaires using homopolar blocks only provided with acceptable reliability and ipsativity when the item bank was properly optimized (i.e.,

minimizing the expected trait estimator variances); 2) the effect of the block selection rules over the recovery of the trait estimates was especially important when using FC CAT assembled *on-the-fly*, being **T**-rule and **A***-rule the ones offering the worst and the best score precision, respectively; and 3) the FC CAT *on-the-fly* offered an improvement in the recovery of trait estimates in comparison with FC CAT using fixed banks.

Limitations and Future Directions

Some limitations of this study are acknowledged. First, as briefly addressed in the Introduction section of this article, the way items are matched in terms of social desirability may have an important effect over the validity of the assessment. In this simulation study, however, the information on the social desirability of the items was not accounted for in the block selection rules. In this sense, recent studies (e.g., Pavlov et al., 2021; Wetzel et al., 2021), have proposed methods for forming blocks based on the distances or agreements between social desirability ratings across items. From the authors' perspective, there is no reason to suspect that the social desirability matching should affect the formation of **A***-optimal FC CATs. In the future, this social desirability constraints may be easily incorporated to the selection rules.

Second, this study focused on the recovery of trait scores with homopolar FC blocks composed of only two items. In this sense, several new studies have been incorporating blocks of more than two items (e.g., Lee & Joo, 2021; Sass et al., 2020; Wetzel et al., 2021; Wetzel & Frick, 2020), as they provide with more bits of information per block used. As pointed out in previous literature (e.g., Brown & Maydeu-Olivares, 2011), increasing the number of items per block should benefit even more the normativity of the trait scores. However, other aspects should be considered. For instance, due to the dependencies between the different pairwise comparisons in each block (e.g., between items 1 and 2, 1 and 3, and 2 and 3), the score reliability was found to be overestimated with blocks of more than two

items (Lin, 2021). Additionally, as found by Sass et al. (2020), subjects report performing pairwise comparisons to respond to FC questionnaires regardless of the number of items per block, thus being the number of pairwise comparisons an indicator of cognitive effort while responding. In this sense, FC questionnaires of pairs and triplets were found to provide with similar reliabilities when the number of pairwise comparisons was the same (i.e., 20 triplets versus 60 pairs, both with 60 pairwise comparisons; Frick et al., 2021). The findings with item pairs in the present article can be regarded as the lower limit performance under the simplest FC format that is used in practice (e.g., Morillo et al., 2019). Although other formats are expected to improve FC CAT performance in terms of true reliability, there is no reason to believe that the effect of the manipulated factors should be different.

Third, due to the size of the block search space, the implementation of FC CAT *on-the-fly* may require a great deal of computational power in order to select the blocks within reasonable time. Although this was not a problem with the simulation conditions included in this study, bigger item pools or larger block sizes may increase the search space exponentially. The block exposure rates presented in Figure 3 of this study indicate that blocks with certain characteristics were rarely used (e.g., with both item discriminations low). Accordingly, future studies may consider investigating how to reduce the search space of FC CAT *on-the-fly*, filtering out the blocks that may not contribute to measure the trait continuum.

Fourth, as previously mentioned, this study assumed item parameters to be invariant between those obtained with single-stimulus responses and those from FC pairs. That it, this assumption implies that the probability of agreement with each individual statement (Equation 3) does not depend on the other statement with which an item is paired. As extensively discussed in the introduction of this article, this may not be always the case. For instance, if the items in a pair strongly differ in social desirability, respondents may tend to select the most

desirable item regardless of their true agreement with each item. Consequently, if this invariance assumption does not hold, using information-based optimization criteria computed with the item parameters estimated from single-stimulus responses may not lead to optimal questionnaires. Therefore, it is of extreme importance to account for the social desirability of the items assembled in the same block. In other words, as a starting point, one should avoid forming blocks that may not be invariant. In this sense, a few approaches for social desirability matching have been proposed recently (e.g., Li et al., 2022; Pavlov et al., 2021; Wetzel & Frick, 2020). Specifically, using social desirability ratings of the items, the requirement of similar social desirability ratings could be set as a constraint in the construction of FC assessments (i.e., non-adaptive questionnaires, fixed block banks, or setting the search space for FC CAT *on-the-fly*). Additionally, the invariance assumption could be empirically tested by recalibrating the blocks with FC responses. To this end, a demonstration on the estimation of the MUPP-2PL for pairwise FC responses using the *mirt* package (Chalmers, 2012) in *R* was made available at https://osf.io/cy5z8, although other software exist (e.g., Brown & Maydeu-Olivares, 2012; Bürkner, 2019). If there is any suspicion that the invariance of some blocks may not hold, practitioners may consider assembling larger questionnaires/block banks using the single-stimulus parameters in order to have a margin for excluding the blocks that might be uninformative in the FC context. As previously pointed out, the assumption of invariance is a key limitation for FC CAT *on-the-fly*, as it may not be feasible to calibrate the complete bank (with every possible block) given the breadth of the search space. As an attempt to provide evidence for the invariance, a subset of the block bank could be calibrated with FC responses.

Finally, the authors chose to investigate the feasibility of using only positively-keyed items, as it represents the most adverse condition in FC questionnaires (see Frick et al., 2021). Although the results found here in this condition were very good, providing with reliable and normative scores under proper optimization criteria, the inclusion of opposite-

keyed items should also improve the reliability and validity of the scores if there are no faking attempts. In this sense, empirical studies are still needed to determine under which assessment conditions the inclusion heteropolar blocks may be appropriate (without affecting the robustness to faking), and when it may not.

References

Brown, A. (2016). Item Response Models for Forced-Choice Questionnaires: A Common Framework. *Psychometrika*, *81*(1), 135–160. https://doi.org/10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2011). Item Response Modeling of Forced-Choice Questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

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 *The Wiley Handbook of Psychometric Testing* (pp. 523–569). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch18

Bunji, K., & Okada, K. (2020). Joint modeling of the two-alternative multidimensional forced-choice personality measurement and its response time by a Thurstonian D-diffusion item response model. *Behavior Research Methods*, *53*(3), 1091–1107. https://doi.org/10.3758/s13428-019-01302-5

Bürkner, P.-C. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software*, *4*(42), 1662.

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

CEB. (2010). *Global personality inventory—Adaptive technical manual.* CEB.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chang, H.-H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methods for the social sciences* (pp. 117–133). SAGE Publications.

Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The Relationship Between Item Exposure and Test Overlap in Computerized Adaptive Testing. *Journal of Educational Measurement*, *40*(2), 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Cheung, M. W.-L., & Chan, W. (2002). Reducing Uniform Response Bias With Ipsative Measurement in Multiple-Group Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(1), 55–77. https://doi.org/10.1207/S15328007SEM0901_4

Clemans, W. V. (1966). *An Analytical and Empirical Examination of Some Properties of Ipsative Measures* (Psychometric Monograph No. 14). Psychometric Society. https://www.psychometricsociety.org/sites/main/files/file-attachments/mn14.pdf

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher's z Transformations. *The Journal of General Psychology*, *125*(3), 245–261. https://doi.org/10.1080/00221309809595548

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personaliry Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Psychological Assessment Resources.

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* (Tech Report No. 1311). U.S. Army Research Institute for the Behavioral and Social Sciences.

Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the Normativity of Trait Estimates from Multidimensional Forced-Choice Data. *Multivariate Behavioral Research*, *0*(0), 1–29. https://doi.org/10.1080/00273171.2021.1938960

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

Houston, J. S., Borman, W. C., Farmer, W. F., & Bearden, R. M. (2006). *Development of the navy computer adaptive personality scales (NCAPS)* (NPRST-TR-06-2). Navy Personnel Research, Studies, and Technology Division, Bureau of Naval Personnel (NPRST/PERS-1).

Joo, S.-H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods*, *52*, 761–772. https://doi.org/10.3758/s13428-019-01274-6

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

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement*, *39*(3), 167–188. https://doi.org/10.1177/0146621614554650

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

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

Li, M., Sun, T., & Zhang, B. (2022). autoFC: An R Package for Automatic Item Pairing in Forced-Choice Test Construction. *Applied Psychological Measurement*, *46*(1), 70–72.

Lin, Y. (2021). Reliability Estimates for IRT-Based Forced-Choice Assessment Scores. *Organizational Research Methods*, 1094428121999086. https://doi.org/10.1177/1094428121999086

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389–414.

Luecht, R. M. (1996). Multidimensional Computerized Adaptive Testing in a Certification or Licensure Context. *Applied Psychological Measurement*, *20*(4), 389–404. https://doi.org/10.1177/014662169602000406

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

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A Silk Purse From the Sow's Ear: Retrieving Normative Information From Multidimensional Forced-Choice Items. *Organizational Research Methods*, *8*(2), 222–248. https://doi.org/10.1177/1094428105275374

McKinley, R. L., & Reckase, M. D. (1982). *The Use of the General Rasch Model with Multidimensional Item Response Data.* AMERICAN COLL TESTING PROGRAM IOWA CITY IA.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*(4), 531–551. https://doi.org/10.1348/0963179042596504

Montano, D., Reeske, A., Franke, F., & Hüffmeier, J. (2017). Leadership, followers' mental health and job performance in organizations: A comprehensive meta-analysis from an occupational health perspective. *Journal of Organizational Behavior*, *38*(3), 327–350. https://doi.org/10.1002/job.2124

Morillo, D. (2018). *Item Response Theory Models for Forced-Choice Questionnaires*.

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. *Revista de Psicología Del Trabajo y de Las Organizaciones*, *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*, 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

Ng, V., Lee, P., Ho, M.-H. R., Kuykendall, L., Stark, S., & Tay, L. (2021). The Development and Validation of a Multidimensional Forced-Choice Format Character Measure:

Testing the Thurstonian IRT Approach. *Journal of Personality Assessment*, *103*(2), 224–237. https://doi.org/10.1080/00223891.2020.1739056

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434. https://doi.org/10.1037/1082-989X.8.4.434

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

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

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

Schulte, N., Holling, H., & Bürkner, P.-C. (2021). Can High-Dimensional Questionnaires Resolve the Ipsativity Issue of Forced-Choice Response Formats? *Educational and Psychological Measurement*, *81*(2), 262–289. https://doi.org/10.1177/0013164420934861

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. https://doi.org/10.1007/BF02294343

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2020). *afex: Analysis of factorial experiments* (R package version 0.28-0). https://CRAN.R-project.org/package=afex

Sorrel, M. A., Abad, F. J., & Nájera, P. (2021). Improving accuracy and usage by correctly selecting: The effects of model selection in cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *45*(2), 112–129.

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

van der Linden, W. J. (1996). Assembling Tests for the Measurement of Multiple Traits. *Applied Psychological Measurement*, *20*(4), 373–388. https://doi.org/10.1177/014662169602000405

van der Linden, W. J. (1999). Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion. *Journal of Educational and Behavioral Statistics*, *24*(4), 398–412. https://doi.org/10.3102/10769986024004398

van der Linden, W. J. (2006). *Linear models for optimal test design*. Springer Science & Business Media.

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*(8), 600–613. https://doi.org/10.1177/0146621617703183

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*, *32*(3), 239–253. https://doi.org/10.1037/pas0000781

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156–170. https://doi.org/10.1037/pas0000971

Tables

Table 1

Trait Correlation Matrix Observed in the NEO PI-R (Costa & McCrae, 1992) *with Neuroticism Reversed to Emotional Stability*

	ES	EX	OE	AG	CO
ES	1				
EX	0.21	1			
OE	0	0.4	1		
AG	0.25	0	0	1	
CO	0.53	0.27	0	0.24	1

Note. ES = emotional stability; EX = extraversion; OE = openness to experiences; AG = agreeableness; CO = conscientiousness.

Table 2

*True reliability and RMSE for Different Assessment Methods, Questionnaire Lengths (J), Selection Rules, and Trait Intercorrelation Matrices (**Φ**).*

		Non-Adaptive Assessment	Adaptive Assessment								
			Random Bank			Optimal Bank			*On-the-fly*		
Φ	*J*		**T**	**D***	**A***	**T**	**D***	**A***	**T**	**D***	**A***
			True Reliability ($r^2_{\theta\hat{\theta}}$)								
Identity	30	0.75	0.77	0.78	0.78	0.80	0.80	0.81	0.74	0.79	0.84
	60	0.84	0.84	0.84	0.84	0.87	0.87	0.88	0.78	0.87	0.91
NEO PI-R	30	0.76	0.73	0.75	0.76	0.78	0.80	0.81	0.63	0.78	0.84
	60	0.84	0.82	0.83	0.83	0.87	0.87	0.88	0.68	0.87	0.91
			Root-Mean-Square Error								
Identity	30	0.50	0.48	0.47	0.47	0.46	0.45	0.44	0.51	0.46	0.41
	60	0.40	0.40	0.40	0.40	0.36	0.36	0.36	0.47	0.37	0.31
NEO PI-R	30	0.50	0.52	0.50	0.49	0.47	0.45	0.44	0.61	0.47	0.41
	60	0.40	0.42	0.42	0.41	0.37	0.36	0.35	0.57	0.37	0.31

Note. **Φ** = true trait correlation matrix; *J* = number of blocks (of two items); **T** = **T**-rule; **D*** = Bayesian **D**-rule; **A*** = Bayesian **A**-rule. The standard deviations of the indicators across replications ranged from 0.004 to 0.015.

Table 3

Eta Square Effect Sizes for the ANOVAs of Trait Estimate Accuracy Indicators

	Non-Adaptive Assessment		Adaptive Assessment					
			Random Bank		Optimal Bank		*On-the-fly*	
	$r^2_{\theta\hat{\theta}}$	$\text{RMSE}_{\hat{\theta}}$	$r^2_{\theta\hat{\theta}}$	$\text{RMSE}_{\hat{\theta}}$	$r^2_{\theta\hat{\theta}}$	$\text{RMSE}_{\hat{\theta}}$	$r^2_{\theta\hat{\theta}}$	$\text{RMSE}_{\hat{\theta}}$
Within-group effects								
Selection Rule	-	-	**0.20***	**0.15***	**0.51***	**0.43***	**0.99***	**0.98***
Selection Rule × *J*	-	-	0.10*	0.06*	**0.17***	0.09*	**0.45***	**0.63***
Selection Rule × **Φ**	-	-	0.03*	0.02*	0.07*	0.04*	**0.91***	**0.85***
Selection Rule × *J* × **Φ**	-	-	0.02*	0.01*	0.02*	0.01*	0.02	0.00
Between-group effects								
J	**0.98***	**0.98***	**0.94***	**0.93***	**0.98***	**0.98***	**0.94***	**0.94***
Φ	0.02	0.02	**0.58***	**0.51***	0.11*	0.06*	**0.84***	**0.74***
J × **Φ**	0.02	0.02	0.08*	0.03	0.05*	0.02	0.02	0.00

Note. *J* = number of blocks (of two items); **Φ** = true trait correlation matrix; $r^2_{\theta\hat{\theta}}$ = true reliability; $\text{RMSE}_{\hat{\theta}}$ = root-mean-square error; * $p < 0.05$. The non-significant interaction effects across all ANOVAs are omitted and large effects (i.e., $\eta^2 \geq 0.14$) are bolded.

Table 4

*Trait Intercorrelation Bias and Average Disattenuated Correlation Between Scores and Criterion for Different Assessment Methods, Questionnaire Lengths (J), Selection Rules, and Trait Intercorrelation Matrices (**Φ**).*

		Non-Adaptive Assessment	Adaptive Assessment								
			Random Bank			Optimal Bank			*On-the-fly*		
Φ	*J*		**T**	**D***	**A***	**T**	**D***	**A***	**T**	**D***	**A***
			Trait Intercorrelation Bias								
Identity	30	-0.07	-0.12	-0.11	-0.10	-0.08	-0.07	-0.06	-0.20	-0.14	-0.05
	60	-0.05	-0.08	-0.08	-0.08	-0.05	-0.04	-0.04	-0.19	-0.09	-0.04
NEO PI-R	30	-0.06	-0.14	-0.12	-0.10	-0.08	-0.06	-0.05	-0.32	-0.13	-0.04
	60	-0.04	-0.09	-0.09	-0.08	-0.05	-0.04	-0.04	-0.30	-0.08	-0.03
			Disattenuated Criterion Validity								
Identity	30	0.23	0.15	0.17	0.19	0.22	0.23	0.25	0.03	0.13	0.25
	60	0.25	0.21	0.21	0.22	0.25	0.26	0.26	0.05	0.20	0.27
NEO PI-R	30	0.31	0.24	0.25	0.27	0.29	0.30	0.31	0.05	0.24	0.30
	60	0.30	0.26	0.27	0.27	0.30	0.30	0.30	0.08	0.27	0.30

Note. **Φ** = true trait correlation matrix; *J* = number of blocks (of two items); **T** = **T**-rule; **D*** = Bayesian **D**-rule; **A*** = Bayesian **A**-rule. The simulated criterion validity was 0.3 for all traits. The standard deviations of the indicators across replications ranged from 0.002 to 0.026.

Table 5

*Item and Block Overlap Rates for Different FC CAT Assembly Methods, Questionnaire Lengths (J), Selection Rules, and Trait Intercorrelation Matrices (**Φ**).*

		Random Bank			Optimal Bank			*On-the-fly*					
		Item/Block Overlap						**Item Overlap**			**Block Overlap**		
Φ	*J*	**T**	**D***	**A***	**T**	**D***	**A***	**T**	**D***	**A***	**T**	**D***	**A***
Identity	30	0.43	0.43	0.44	0.39	0.40	0.40	0.73	0.71	0.64	0.04	0.03	0.03
	60	0.61	0.61	0.63	0.56	0.57	0.60	0.81	0.81	0.78	0.03	0.02	0.02
NEO PI-R	30	0.45	0.45	0.47	0.40	0.40	0.43	0.74	0.70	0.64	0.05	0.03	0.03
	60	0.63	0.63	0.65	0.57	0.58	0.62	0.82	0.82	0.78	0.04	0.03	0.03

Note. **Φ** = true trait correlation matrix; *J* = number of blocks (of two items); **T** = **T**-rule; **D*** = Bayesian **D**-rule; **A*** = Bayesian **A**-rule. The standard deviations of the indicators across replications ranged from 0.001 to 0.017.

Figures

Figure 1

Schematic of the Adaptive Algorithm.

1. Randomly administer three blocks addressing the 5 dimensions measured (e.g., 1-2, 3-4, 4-5).
2. Remove from the bank every block that includes the items already administered.
3. If the maximum number of blocks for a given dimension pair is reached (3 and 6 for J = 30 and 60, respectively), remove from the bank all blocks measuring this pair of traits.
4. Calculate $\widehat{\boldsymbol{\theta}}$ with Expected-A-Posteriori method.
5. Select the block that best optimizes the criterion ($\mathbf{T}$, $\mathbf{D}^*$ or $\mathbf{A}^*$) for the current $\widehat{\boldsymbol{\theta}}$.
6. Remove from the bank every block that includes the items already administered.
7. If the maximum number of blocks for a given dimension pair is reached, remove from the bank all blocks measuring this pair of traits.
8. Repeat Steps 4 to 7 until FC CAT length J is achieved.
9. Calculate final $\widehat{\boldsymbol{\theta}}$.

Figure 2

*Average Conditional Standard Errors of Estimates for Different Assessment Methods and Trait Intercorrelation Matrices (**Φ**) using **A***-Rule and 60 blocks.*

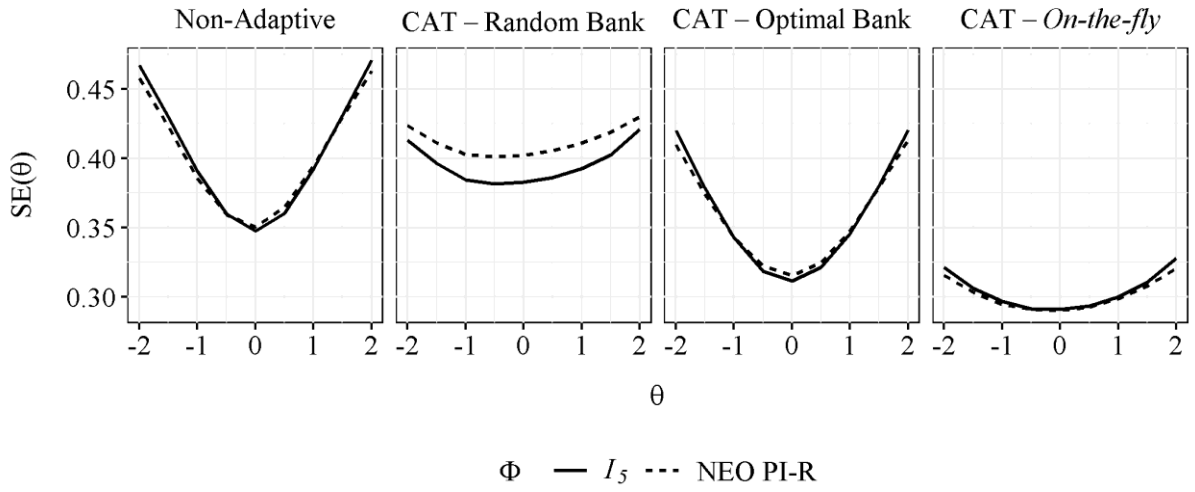

Figure 3

Distribution of Scale Parameters in Pairs for Different Selection Rules Under Assessments On-the-fly (with one replication with 60 blocks).