METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

A Comparison of Optimization Algorithms for Forced-Choice Questionnaire Assembly

Scarlett Escudero [1,2], Miguel A. Sorrel [2], Rodrigo S. Kreitchmann [3], Francisco J. Abad [2]

[1] *Department of Educational Psychology, University of Minnesota, Minneapolis, MN, USA.* **[2]** *Department of Social Psychology and Methodology, Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain.* **[3]** *Department of Methodology of Behavioral Sciences, Faculty of Psychology, Universidad Nacional de Educación a Distancia, Madrid, Spain.*

Methodology, 2026, Vol. 22(2), 172–194, https://doi.org/10.5964/meth.18925

Received: 2025-07-18 • **Accepted:** 2026-03-06 • **Published (VoR):** 2026-06-30

Handling Editor: Pablo Nájera Álvarez, Universidad Pontificia Comillas, Madrid, Spain

Corresponding Author: Miguel A. Sorrel, Faculty of Psychology, Universidad Autónoma de Madrid, 6 Iván Pavlov St, Cantoblanco Campus, Madrid, Spain, 28049. E-mail: miguel.sorrel@uam.es

Supplementary Materials: Code, Data, Materials [see Index of Supplementary Materials]

Abstract

Forced-choice questionnaires (FCQs) are increasingly favored over traditional Likert-type formats due to their reduced susceptibility to faking and social desirability (SD). Their construction typically involves pairing items from existing single-stimulus banks. This study compares four methods for assembling FCQs: a genetic algorithm (GA), two simulated annealing (SA) strategies (blueprint-based and scale-parameter-optimized), and brute-force (BF) random search. These methods are evaluated via simulation and an empirical example, focusing on trait score recovery. The effects of questionnaire length and SD matching on recovery are also examined. Three item banks varying in the a_j-SD_j relationship and inclusion of heteropolar blocks were used to assess performance across pairing scenarios. GA consistently produced the most reliable scores, followed by SA with a_j optimization. All examined factors significantly affected reliability. GA is recommended for FCQ assembly, especially with short questionnaires, no heteropolar blocks, and high a_j-SD_j correlation.

Keywords

forced-choice questionnaires, reliability, optimal assembly, genetic algorithm, simulated annealing algorithm, brute-force

Non-cognitive constructs have long been of interest in psychological research and have traditionally been measured using single-stimulus formats (SS), such as the Likert format. However, this format has been known to suffer from response biases such as acquiescence and faking, which can undermine reliability, validity, and variability of the observed scores (Salgado, 2016), and alter the item covariance structure (McCrae et al., 2001). A solution that has shown promise in solving some of these problems is the forced-choice (FC) format. This format increases criterion-related validity (Salgado & Táuriz, 2014) and reduces both faking (Cao & Drasgow, 2019) and other response biases (Kreitchmann et al., 2019). However, recent studies have emphasized that these advantages are only achievable if forced-choice questionnaires (FCQs) are carefully designed (Graña et al., 2025).

Given the numerous possible item combinations involved in block construction, several procedures have been developed to optimize the assembly process. However, to date, a comprehensive comparison of these methods is lacking. The present study aims to address this gap by evaluating the most effective method for assembling FCQs from a SS item bank. The comparison is limited to the two options available in the software at the time of writing, namely the genetic algorithm (GA; Kreitchmann et al., 2022) and the simulated annealing algorithm (SA; Li et al., 2022). While a linear programming approach is possible, it can be substantially more computationally demanding and should be explored in detail in future research. Other heuristics, such as ant colony optimization, have not yet been applied. Therefore, a systematic comparison of available methods is needed to guide researchers in assembling FCQs, especially since GA is implemented in a Shiny app[1] and SA is available in the *autoFC* R package.[2]

Forced-Choice Questionnaire Design and Modeling

In recent years, there has been a growing trend toward the development and use of FCQs to assess non-cognitive constructs (Lee et al., 2025). The FC format can be distinguished from SS format in that a choice must be made among the alternatives rather than rating each statement. A commonly used example is a FC block consisting of item pairs, in which two SS items are presented together and the respondent is asked to choose one over the other (see Figure 1; for a comprehensive review, see Hontangas et al., 2015).

Despite the advantages of FCQs over SS, this response format can introduce fully or partially ipsative scores. Ipsativity refers to the interdependence among trait scores, meaning that if a person scores higher on one trait, they must score lower on another. This can affect validity. For example, in a purely ipsative FCQ, the validity coefficients of all measured traits with respect to a given external criterion will sum (and average)

1) https://psychometricmodelling.shinyapps.io/FCoptimization/

2) https://cran.r-project.org/web/packages/autoFC/index.html

Figure 1

Examples of Non-Cognitive Questionnaire Formats

(Likert-type) Select the option that best describes you.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I consider myself a responsible person.	◯	◯	◯	◯	◯

(FC) Select the statement that best describes you.

I consider myself a responsible person.	I tend to enjoy being around other people.
◯	◯

to zero (Hicks, 1970). Full rank of the effective FCQ loading structure is a necessary condition for achieving non-ipsative scores, as noted in Brown (2016). Within the Thurstonian item response theory (IRT) framework, this condition is typically expressed in terms of the factor loading matrix. For instance, the FCQ loading matrix becomes rank-deficient when the loadings within every block or within every dimension are equal. A straightforward way to avoid this issue is to combine items with factor loadings of opposite sign within each block. Under alternative IRT parameterizations, however, the same rank condition can be expressed in terms of item scale parameters rather than factor loadings. Morillo (2018, p. 74) further illustrates that under such parameterizations, the matrix can also become rank-deficient when the scale parameters of the two dimensions represented in a block maintain a constant ratio across all blocks measuring the same pair of dimensions. While these conditions are violated in classical test theory scoring using blocks of equally keyed items (with equal weights for all items), under IRT modeling, where scale parameters are allowed to vary, such violations occur only on rare occasions. Some available IRT models allow researchers formally characterize the response processes underlying FC formats and obtain non-ipsative scores (e.g., Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; Stark et al., 2005).

Ipsativity is therefore a property of the scoring method rather than of the FC format itself. Hence, ipsativity can be addressed with models such as the Multi-Unidimensional Pairwise Preference Two-Parameter Logistic (MUPP-2PL; Morillo et al., 2016) or the Thurstonian IRT for FC data (TIRT; Brown & Maydeu-Olivares, 2011). For binary FC comparisons, these models yield nearly equivalent response probabilities, although differences increase for block sizes greater than two. Test assembly assumes that item parameters are invariant when moving from SS to pairwise FC administrations. Empirical evidence supports approximate invariance: Lin and Brown (2017) found TIRT parameters largely stable across block compositions (quads vs. triplets), while Morillo et al. (2019)

reported that MUPP-2PL parameters from FC blocks closely matched those from graded-scale formats. In this study, we adopt the MUPP-2PL framework, for which invariance between Likert-type and FC formats has been directly evaluated, showing correlations above .90 across formats.

Multi-Unidimensional Pairwise Preference Two-Parameter Logistic Model

The MUPP-2PL model can be used in dichotomous FC blocks, where the probability of agreement with response option is modeled following a 2PL function. The model includes an invariance assumption, which states that item parameters remain constant regardless of the response format (i.e., FC vs. Likert) and the within-block context (which item is paired with another). Therefore, a_{j_1} and a_{j_2} in Equation (1) should be the same for 2PL items and in the FC block. The block characteristic function is:

$$P\left(Y_{ij}=1\middle|\theta_i\right)=\frac{1}{1+e^{-\left(a_{j_1}\theta_{\widetilde{j_1}i}-a_{j_2}\theta_{\widetilde{j_2}i}+d_j\right)}}, \tag{1}$$

$$d_j=a_{j_2}b_{j_2}-a_{j_1}b_{j_1}, \tag{2}$$

where $Y_{ij}=1$ indicates that the respondent selected item j_1 (the first item in the pair); θ_i represents a person's position on each of the D latent traits measured by the FCQ; $\theta_{\widetilde{j_1}i}$ and $\theta_{\widetilde{j_2}i}$ are the coordinates of θ_i for items j_1 and j_2; a_{j_1} and a_{j_2} are the scale parameters; d_j is the block intercept parameter; and b_{j_1} and $b_{j_2}$are the location parameters. The sign of the scale (discrimination) parameter determines the item's polarity. A positive scale parameter signifies a positively keyed item, where higher trait levels are associated with greater agreement. Conversely, a negative scale parameter indicates a negatively keyed item, where higher trait levels correspond to lower agreement. If both items have the same sign, either positive or negative, the block is considered homopolar or equally keyed; otherwise, it is called heteropolar or unequally keyed. In any case, it must be ensured that the resulting factor loading matrix is of full rank so that non-ipsative scores can be obtained.

Forced-Choice Questionnaire Optimal Assembly

One common approach for designing a FCQ is to pair items from a SS item bank to form blocks. This process can yield thousands of potential FCQs with varying levels of reliability. Because reliability is essential for validity, assembling blocks without considering item properties may result in suboptimal tests. With the growing use of FCQs, the need for optimal assembly procedures has become increasingly important.

Social Desirability

One important consideration in the design of FCQs is the control of social desirability (SD). If SD matching (SDM) is not carefully considered during block construction, differences in item SDs are likely to emerge. This is especially true for heteropolar pairs. Such differences may compromise the validity of the questionnaire scores by making it easier for respondents to select socially desirable options instead of those that reflect their true traits. Matching items by their level of SD is a widely recommended strategy for reducing faking in FCQs (Cao & Drasgow, 2019; Pavlov et al., 2021).

One approach for matching items based on SD involves convening an expert committee to rate the desirability level of each item. The average desirability rating is then calculated for each item across all raters and items are matched by minimizing the absolute difference between items' mean desirability values. Items are considered well-matched when the difference in SD falls below a predetermined cutoff. For homopolar blocks, a common cutoff is 0.50 on a 5-point scale, whereas for heteropolar blocks, the cutoff typically needs to be relaxed in order to ensure that a sufficient number of valid heteropolar pairs can be formed (Graña et al., 2025; Li et al., 2025).

Scale Parameters

The ability to obtain normative scoring from FCQ responses through IRT relies on the differential weighting of responses (Bürkner, 2022), which is driven by the scale parameters. Heteropolar blocks naturally lead to this differential weighting. However, this is a topic of debate in FCQ design since pairing them can be tricky and increase cognitive demand. On the one hand, studies (e.g., Brown & Maydeu-Olivares, 2011; Frick et al., 2023) suggest that including heteropolar blocks can enhance estimation accuracy and validity. On the other hand, recent empirical findings (Graña et al., 2025) question whether heteropolar blocks are actually necessary. These authors suggest that if FCQs are assembled with homopolar blocks that differ in scale parameters, heteropolar blocks are not necessary. This is an ongoing debate, with many authors suggesting that including approximately 15–40% heteropolar blocks can enhance the psychometric properties of FCQs, even if it means accepting a minor compromise in SD (Lee et al., 2022; Li et al., 2025). Thus, achieving faking resistance requires careful SDM, especially when heteropolar blocks are included.

Algorithms for Optimal Assembly

Simulated Annealing Algorithm — The simulated annealing algorithm is a heuristic optimization method inspired by the physical annealing of solids. It operates in two steps (Kirkpatrick et al., 1983): first, the system's temperature is raised to a maximum, then gradually decreased until a minimum is reached, minimizing the system's energy, which corresponds to the cost of a solution. A key feature of SA is its ability to accept worse solutions to escape local optima. Li et al. (2022) implemented SA to assemble

FCQs in the *autoFC* R package. The procedure begins with a user-defined blueprint specifying the number of blocks, block size, trait composition, and keying constraints. Each item is characterized by numerical attributes, such as SD or factor loadings, and the algorithm computes a weighted composite energy for each block, combining block-level indices for each attribute using user-specified weights so that higher absolute values reflect more desirable configurations. Heteropolar blocks are not included in the energy calculation; instead, the user defines specific trait combinations, and blocks meeting these conditions are randomly selected. Starting from a random admissible assembly, the algorithm iteratively swaps or replaces items, accepting changes that lower energy while occasionally allowing higher-energy solutions early on to avoid local optima. As the temperature decreases, the search converges on the lowest-energy arrangement, yielding FC blocks that satisfy the psychometric constraints. For a detailed tutorial, see Li et al. (2025). Version 0.2.0.1002 of the *autoFC* package was used in this study.

Genetic Algorithm — Genetic algorithms are heuristic optimization methods inspired by principles of population genetics. Kreitchmann et al. (2022) adapted the GA for FCQ assembly to maximize the marginal reliability of selected blocks. In this approach, new candidate blocks are generated using a node histogram-based sampling algorithm, which constructs probabilistic models from the genotypes of previous generations. Each new genotype is formed in two steps: first, a portion of the parent genotype is directly passed to the offspring as a template; second, the remaining elements are sampled from a conditional probability distribution capturing dependencies observed in prior generations, with a mutation factor added as noise. Candidates are then evaluated against their parents based on constraint compliance and the objective function, which is to maximize block reliability, and the best candidates advance to the next generation. Block content constraints are represented in a JxJ matrix (**C**), where each cell is 1 if a pair of items can be combined into a block based on content criteria and 0 if not. The node histogram records the frequency of item selection within a generation. This iterative process continues until convergence, producing FCQ blocks that satisfy psychometric constraints and maximize reliability. More details on the algorithm can be found in Kreitchmann et al. (2022).

The Present Study

To date, there has been no comprehensive evaluation of existing methods for the optimal assembly of FCQs. The purpose of this study is to compare such methods: the GA, the SA algorithm, in two approaches, one as blueprint and the other optimizing the a_j parameter differences, and a brute-force (BF) search, focusing on the reliability of the assembled questionnaires' scores and computational cost. The two SA variants differ in their optimization criterion; the former employs the basic functionality of the package, incorporating only content and SDM constraints, without considering item parameters,

whereas the latter further defines an additional criterion that seeks to maximize differences in the scale parameters within each item pair. This study makes a novel contribution by incorporating SD constraints into GA and exploring how its relationship with item parameters may affect the block assembly process. We conducted a simulation study evaluating the four methods and an empirical illustration. We hypothesize that: (1) GA will perform best in terms of reliability, as it incorporates the reliability of the assembled blocks as the objective function to optimize, and be followed by the SA with scale parameter, a_j, optimization method, SA blueprint, and BF; (2) there should be no significant differences in the inclusion or exclusion of heteropolar blocks; (3) a higher correlation of a_j with SD_j is expected to reduce performance; and (4) SDM will negatively affect trait recovery. These hypotheses were not preregistered.

Method

Simulation Design and Data Generation

Three factors were systematically manipulated in the simulation study. For clarity, we categorize these factors into two groups, block factors, which relate to the FCQ construction, and an item factor, which pertains to the construction of the item banks. The block factors are: 1) questionnaire length (40, 80), and 2) use of SDM (Yes, No). The item factor is: 3) degree of correlation between scale parameter, a_j, and SD_j of the positively keyed bank, which relates to the types of blocks formed (homopolar vs. heteropolar), and results in four levels ($r_{a_j^+ - SD_j^+} = 0.20$ with 25% heteropolar blocks, $r_{a_j^+ - SD_j^+} = 0.20$ with 0% heteropolar blocks, $r_{a_j^+ - SD_j^+} = 0.50$ with 0% heteropolar blocks, $r_{a_j^+ - SD_j^+} = 0.80$ with 0% heteropolar blocks). Hereafter, "+" indicates items that are positively keyed to the trait, and "-" indicates items that are negatively keyed. All factors are fully crossed, resulting in 16 conditions.

First, the questionnaire length factor determines how many pairwise blocks are formed. We established two levels, 80 and 40 blocks. Second, the SDM factor indicates whether item pairs were matched based on their SD ratings, with two levels, yes (matching applied) and no (matching not applied). When SDM is applied, the absolute difference between two items' SD_j ratings was calculated; if this difference exceeded a predefined cutoff, the items were not eligible for pairing. We established a cutoff of 0.5 for homopolar blocks to ensure a stricter level of matching. However, since it is more difficult to find heteropolar blocks that match in SD (Graña et al., 2025), the cutoff was relaxed to 0.75 for heteropolar blocks.

Since the third factor pertains to the generation of the item banks, we will describe the item bank generation process together. One five-dimensional bank of 320 SS items was generated for each condition and replication, as to imitate personality item pools, such as the International Personality Item Pool (IPIP; Goldberg, 1999). Three prototypical

item banks were created. The main difference among them lies in the degree of association between the scale parameter, a_j, and SD_j of the positively keyed items. All banks were balanced with 64 items per trait; in the mixed keyed bank, each trait had 32 positive and negative items. Each simulation condition used a separate item bank, corresponding to one of the three bank types described below.

The item banks differed in the degree of correlation between a_j and SD_j, as well as in block polarity, resulting in four distinct categories (Table 1). Bank 1 was used for Categories 1 and 2 (and Levels 1 and 2 of the third simulation factor). It included both positively keyed items and negatively keyed items, with a small correlation between a_j and SD_j for both types ($r_{a_j^+ - SD_j^+} = r_{a_j^- - SD_j^-} = 0.20$) and a naturally high correlation across keys ($r_{a_j - SD_j} = 0.86$). Using this bank, Category 1 formed a FCQ with 25% heteropolar blocks, while Category 2 included only homopolar blocks (0% heteropolar).

Bank 1, which reflects a more realistic scenario, is expected to pose less difficulty in assembling reliable tests due to the low correlation between a_j and SD_j. To explore more challenging conditions, we included two additional item banks. Levels 3 and 4 of the third simulation factor (and Categories 3 and 4) correspond to Banks 2 and 3, respectively; both consist of positively keyed items and therefore can form only homopolar blocks. Bank 2 was generated with a moderate correlation between the scale parameter, a_j, and SD_j ($r_{a_j^+ - SD_j^+} = 0.50$), while Bank 3 has a high correlation ($r_{a_j^+ - SD_j^+} = 0.80$), making optimal block matching more difficult. The SD_j values were sampled this way to represent that, in real contexts, when traits are scored in the socially desirable direction (i.e., conscientiousness, emotional stability), positively keyed items tend to have higher desirability. The values in these distributions were primarily based on the empirical distributions of a_j, d_j, and SD_j reported in the publicly available datasets from Johnson (2014) and Hughes et al. (2021), while also considering that positively keyed items tend to exhibit higher SD values than negatively keyed items (Graña et al., 2025; Li et al., 2025). Under our scoring convention, positively keyed items for each trait were defined as the socially desirable direction (e.g., for items measuring Neuroticism, a positively keyed item has a lower SD rating). These choices were made to represent realistic item parameters and SD behavior. This also includes the correlation between a_j and d_j (approximately 0.30 for both positively and negatively keyed items). The correlation between d_j and SD_j was kept at zero so that SD_j would only be linked to a_j, allowing us to analyze the impact of this variable.

Methodology
2026, Vol. 22(2), 172–194
https://doi.org/10.5964/meth.18925

Table 1

Parameter Simulation Specifications

	Bank 1		Bank 2	Bank 3
Parameter	**(+)**	**(-)**	**(+)**	**(+)**
a_j	N(1.5,0.5)	N(-1.5,0.5)	N(1.5,0.5) truncated at 0	N(1.5,0.5) truncated at 0
d_j	N(-0.5,0.8) truncated at -3	N(-1,0.8) truncated at -3	N(-0.5,0.8)	N(-0.5,0.8)
SD_j	N(4,0.5) truncated at 1 and 5	N(2,0.5) truncated at 1 and 5	N(4,0.5) truncated at 1 and 5	N(4,0.5) truncated at 1 and 5
$r_{a_j - SD_j}$	.20	.20	.50	.80
$r_{a_j - d_j}$	.30	-.30	.30	.30
$r_{d_j - SD_j}$	.00	.00	.00	.00

Note. a_j = scale parameter; d_j = block intercept parameter; SD_j = social desirability of each item; $r_{a_j - SD_j}$ = correlation between a_j and SD_j; $r_{a_j - d_j}$ = correlation between a_j and d_j; $r_{d_j - SD_j}$ = correlation between d_j and SD_j.

The structure of the simulation is as follows. First, a SS item bank was generated for each condition and replication. Second, a FCQ was constructed with each algorithm using the SS parameters. Then, a binary FC response dataset of 5,000 respondents was simulated using the MUPP-2PL for each assembled FCQ. Finally, the MUPP-2PL was estimated, and trait recovery was assessed for each questionnaire. Trait estimates ($\widehat{\theta}$) were obtained as maximum a posteriori scores with the Metropolis-Hastings Robbins-Monro algorithm. These analyses were conducted using the *mirt* R package (Chalmers, 2012). This process was replicated 50 times. We controlled for the following content constraints: (1) block multidimensionality (i.e., each block had to include items measuring two different traits), and (2) trait balance across the selected blocks (i.e., each trait had to be represented by the same number of items). Item repetition was not allowed. During FCQ assembly, we verified that content and polarity constraints were met and that the distribution of blocks and items across traits remained generally balanced, allowing a fair comparison of the algorithms. Additionally, we conducted simulation checks to evaluate if any of the resulting scale parameter matrices after FCQ assembly were rank restricted by analyzing the least singular value of such matrices. Across all conditions, assembly algorithms, and replications, the least singular value was strictly greater than zero, indicating that none of the matrices were rank-deficient. We recorded the time spent in seconds for each assembly algorithm. In the case of BF, 100 questionnaires that met the specified constraints and SD requirements were randomly formed and the one with the highest reliability was selected. The R code used for the analysis and the empirical study can be found at Sorrel et al. (2026). The repository also includes a document detailing all algorithm specifications. All procedures were executed with a 2.50 GHz Intel Core i9-11900 CPU and 32 GB of RAM.

Measures of Trait Recovery

To compare the assembly methods, the main dependent variable was trait score recovery. Specifically, we computed for each replica and trait: (1) the true reliability, calculated using the squared correlation between estimated and true θ ($\rho^2_{\widehat{\theta}\theta}$); (2) the root mean square error of $\widehat{\theta}$, both overall ($RMSE_{\widehat{\theta}}$; Equation 3) and conditional to the true θ ($RMSE_{\widehat{\theta}}(\theta)$). The conditional $RMSE_{\widehat{\theta}}$ was calculated within intervals of the true θ to examine how estimation accuracy varies across the latent continuum. Individuals were grouped into bins of 0.5 spanning from -2 to 2, and within each bin the RMSE was computed from the squared estimation errors of all individuals in that group. Additionally, an indicator of ipsativity was included, consisting of (3) the average trait correlation bias ($Bias_{\widehat{\Phi}}$; Equation 4).

$$RMSE_{\widehat{\theta}} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\theta}_s - \theta_s\right)^2} \tag{3}$$

$$\text{Bias}_{\widehat{\Phi}} = \widehat{\Phi} - \Phi, \tag{4}$$

where $\widehat{\Phi}$ and Φ are the estimated and true trait correlation matrices, respectively. We used the real-world correlations from the NEO-PI-R (Costa & McCrae, 1992). In the case of fully ipsative scores, a negative bias of $-1/(D-1)$ would be expected for $\widehat{\Phi}$, with D representing the number of traits (Hicks, 1970). Dependent Variables (1) and (2) were calculated separately for each trait and then averaged across all five traits, whereas (3) was calculated by extracting the non-diagonal elements of $\widehat{\Phi}$ and Φ, applying Fisher's Z-transformation to each correlation, computing the Z-differences, averaging these differences, and then back-transforming the average to the correlation metric (Corey et al., 1998). Results of the overall $\text{RMSE}_{\hat{\theta}}$ and of the four-way univariate analyses of variance (ANOVA) for each dependent variable, where algorithm was treated as a within-condition factor and the simulation conditions as between-condition factors, can be found in Tables S2 and S3 in the Supplementary Material (see Escudero et al., 2026). Partial eta-squared (η_p^2) values higher than .14 were considered as relevant effects (Cohen, 1988). Results for $\text{RMSE}_{\hat{\theta}}$ (shown in Figure S1, Escudero et al., 2026) are omitted from the main text, as the conclusions are the same as for $\rho^2_{\hat{\theta}\theta}$. Therefore, we focus on the latter, which is a more commonly used metric. ANOVA results were used to guide the interpretation of the findings.

Results

Algorithm Efficiency

The GA is notably influenced by the questionnaire length, with longer questionnaires resulting in increased duration, as shown in Table S1 in the Supplementary Material (see Escudero et al., 2026). Assembling questionnaires of 80 and 40 blocks took an average of 4.90 and 3.21 minutes, respectively. In contrast, the other algorithms show minimal sensitivity to questionnaire length. Both SA methods are typically completed in a few seconds. However, SA has higher skewness and kurtosis, as we implemented an iterative process that reruns the algorithm until the target design is achieved. Across all conditions and replications, the constraint on the number of heteropolar blocks was always satisfied. In a small number of replications of the SA method, however, a few blocks did not meet the SDM constraint. Out of the 400 replications with SDM, there were 28 with 1 block affected and 2 with 2 blocks affected for the SA blueprint, and SA with Scale Parameter a_j optimization showed 54 with 1 block affected and 1 with 2 blocks affected. The small number of affected blocks (1 or 2 out of 40 or 80) suggests that the impact is negligible. In any case, this implies a potential advantage in reliability and a disadvantage in ipsativity for these replications.

Recovery of Trait Parameters

Table 2 presents the marginal results for the $\rho^2_{\hat{\theta}\theta}$ and $\text{Bias}_{\hat{\Phi}}$ of the four assembly methods. GA consistently yields the highest reliability ($M = .80$), followed by SA with a_j optimization ($M = .78$), BF ($M = .76$), and SA blueprint ($M = .73$). These differences indicate a large effect size ($\eta^2_p = .88$). All simulation conditions are relevant, as seen in Figure 2. The most relevant factor was the questionnaire length, which can be expected, as longer questionnaires enhance reliability ($\eta^2_p = .98$). Shorter tests show lower reliability overall, with the same relative patterns across correlations, heteropolar proportions, and algorithms. For instance, controlling SD at $r_{a_j^+ - SD_j^+} = .80$ with 0% heteropolar blocks, GA's reliability is .71 for Length 40, compared to .80 for Length 80. The next relevant factor was the degree of relation between the scale parameter, a_j, and SD_j in the item bank and having heteropolar blocks or not ($\eta^2_p = .95$). The highest reliability is achieved with mixed banks forming questionnaires containing 25% heteropolar blocks; for example, for Length 80, controlling for SD at $r_{a_j^+ - SD_j^+} = .20$, GA achieves .88 vs. .84 in the same condition with 0% heteropolar blocks. As the correlation between the scale parameter and SD_j increases, reliability decreases. The same tendency can be seen in both test lengths. Specifically, the condition with the lowest reliability results is the positively keyed item bank with $r_{a_j^+ - SD_j^+} = .80$ in the 40-block test length and it is where the most differences between algorithms are seen: GA = .71, $\text{SA}_{\text{a}_\text{j}}$ = .68, BF = .65, and SA_{bp} = .60. Although the use of SDM had the smallest effect size among the factors examined ($\eta^2_p = .53$), it still had a meaningful impact on reliability, consistently leading to lower values, due to the constraints it imposes on item pairing. Specifically, taking GA as an example, reliability was consistently lower when incorporating the desirability constraint, decreasing on average by .01 points compared to the same condition without considering SD, and reaching a decrease of up to .04 points in cases where there is a strong relationship between the scale parameter and SD_j. The same pattern was observed for the other procedures. Notable interactions were; algorithm with length ($\eta^2_p = .31$), algorithm with D-H ($\eta^2_p = .54$) and SDM with D-H ($\eta^2_p = .59$). These interactions further prove that these factors are relevant in assembling FCQs, as the algorithm proves to be more important when the correlation between a_j and SD_j is higher and whether you form heteropolar blocks or not, and that this same factor interacts with SDM. In all cases, the ordering of the methods described above is preserved.

Table 2

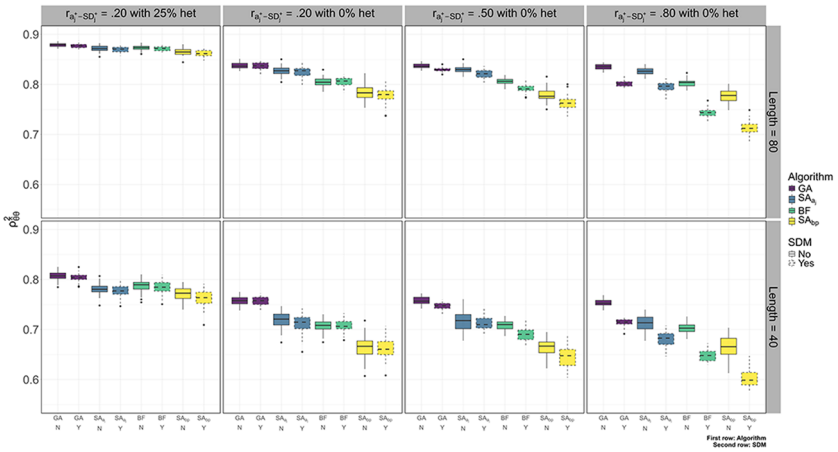
Average Trait Recovery Across 50 Replications for Questionnaires Assembled Using Each Algorithm in the Simulation Study

			$\rho^2_{\theta\theta}$				$\text{Bias}_{\hat{\Phi}}$			
Length	**D-H**	**SDM**	**GA**	**SA$_{a_j}$**	**BF**	**SA$_{b_p}$**	**GA**	**SA$_{a_j}$**	**BF**	**SA$_{b_p}$**
80	$r_{a_j^+ - SD_j^+}$ = .20 with 25% het	Yes	**.88**	.87	.87	.86	.01	.01	.01	**.00**
		No	**.88**	.87	.87	.86	**.01**	**.01**	**.01**	**.01**
	$r_{a_j^+ - SD_j^+}$ = .20 with 0% het	Yes	**.84**	.82	.81	.78	**-.05**	-.07	-.10	-.14
		No	**.84**	.83	.81	.78	**-.05**	-.06	-.10	-.13
	$r_{a_j^+ - SD_j^+}$ = .50 with 0% het	Yes	**.83**	.82	.79	.76	**-.06**	-.07	-.12	-.17
		No	**.84**	.83	.81	.78	**-.05**	-.06	-.09	-.14
	$r_{a_j^+ - SD_j^+}$ = .80 with 0% het	Yes	**.80**	**.80**	.74	.71	**-.10**	-.11	-.21	-.26
		No	**.83**	.83	.80	.78	**-.05**	-.06	-.10	-.14
40	$r_{a_j^+ - SD_j^+}$ = .20 with 25% het	Yes	**.80**	.78	.78	.76	.01	.01	**.00**	**.00**
		No	**.81**	.78	.79	.77	.02	.01	.01	**.00**
	$r_{a_j^+ - SD_j^+}$ = .20 with 0% het	Yes	**.76**	.71	.71	.66	**-.06**	-.12	-.14	-.20
		No	**.76**	.72	.71	.66	**-.05**	-.11	-.13	-.21
	$r_{a_j^+ - SD_j^+}$ = .50 with 0% het	Yes	**.75**	.71	.69	.64	**-.07**	-.12	-.17	-.24
		No	**.76**	.72	.71	.66	**-.05**	-.12	-.13	-.20
	$r_{a_j^+ - SD_j^+}$ = .80 with 0% het	Yes	**.71**	.68	.65	.60	**-.12**	-.17	-.25	-.32
		No	**.75**	.71	.70	.66	**-.06**	-.12	-.14	-.21
Grand mean			**.80**	.78	.76	.73	**-.04**	-.07	-.10	-.15

Note. Maximum values of $\rho^2_{\theta\theta}$ and $\text{Bias}_{\hat{\Phi}}$ closest to zero are marked in bold. All standard deviations of $\rho^2_{\theta\theta}$ and $\text{Bias}_{\hat{\Phi}}$ range around 0 and .03. D-H = degree of relation between a_j and SD_j of the positively keyed item bank and possibility of forming heteropolar blocks; SDM = social desirability matching; GA = genetic algorithm; SA$_{a_j}$= simulated annealing with scale parameter, a_j, optimization; BF = brute-force; SA$_{b_p}$ = simulated annealing blueprint.

Figure 2

Squared Correlation Between $\hat{\theta}$ *and* θ



Note. GA = genetic algorithm; SA_{a_j}= simulated annealing with scale parameter, a_j, optimization; BF = brute-force; SA_{bp} = simulated annealing blueprint; SDM = social desirability matching. Within every facet, the four algorithms are always displayed in the same left-to-right order.

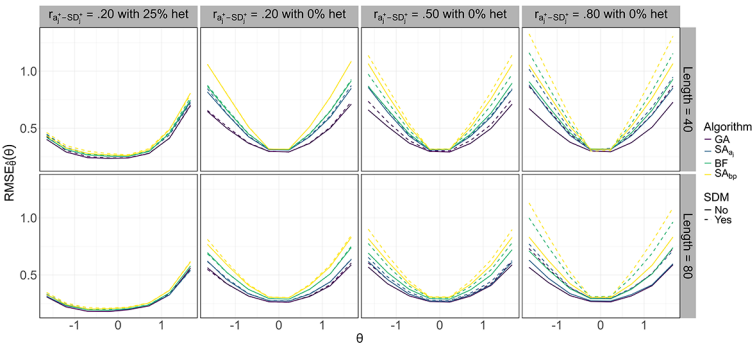
As for $RMSE_{\hat{\theta}}(\theta)$, Figure 3 shows relevant differences across algorithms that align with the results found in $\rho^2_{\hat{\theta}\theta}$. Generally, GA has lesser error, as it is closer to zero, and SA blueprint is the furthest method from zero. RMSE is lower for θ values near zero and increases at the extremes, as well as longer questionnaires reduce error, and greater differences among the algorithms are observed in the condition where $r_{a_j^+ - SD_j^+} = .80$ in the 40-block questionnaire.

Ipsativity Indicator

Regarding the ipsativity indicator ($Bias_{\hat{\Phi}}$), several patterns emerge from the absolute values in Table 2. As seen in Table S3 in the Supplementary Material (see Escudero et al., 2026), the ANOVA results follow the same pattern as the $\rho^2_{\hat{\theta}\theta}$ results previously described. First, mixed banks with 25% heteropolar blocks produce negligible bias across all algorithms ($Bias_{\hat{\Phi}}$ ranges from .00 to .01). In contrast, bias increases as the correlation between the scale parameter, a_j, and SD_j grows and when only positively keyed items are used. For the most demanding condition (i.e., $r_{a_j^+ - SD_j^+} = .80$, 0% heteropolar blocks, SDM, and Length 40), the bias reaches -.12 for GA, -.17 for SA with a_j optimization, -.25 for BF, and -.32 for SA blueprint. Across all conditions, GA consistently shows the lowest ipsativity levels, followed by SA with a_j optimization, BF, and SA blueprint. This ordering is also reflected in the grand means: -.04 (GA), -.07 (SA_{a_j}), -.10 (BF), and -.15 (SA_{bp}). These differences also have a large effect size ($\eta^2_p = .82$). Thus, while all methods show

Figure 3

Average Conditional Root Mean Square Error



Note. GA = genetic algorithm; SA_{a_j}= simulated annealing with scale parameter, a_j, optimization; BF = brute-force; SA_{bp} = simulated annealing blueprint; SDM = social desirability matching.

increasing bias, as the conditions are more demanding, GA systematically yields the least distorted correlation matrices.

Empirical Illustration

We applied GA, SA, both the blueprint method and with a scale parameter, a_j, optimization, and BF to an empirical dataset consisting of 286 Likert-type mixed item bank (145 positively keyed) from the IPIP-NEO (Johnson, 2014). From this dataset, we drew a random sample of 1,000 U.S. participants aged between 19 and 25 years old with no missing data. The rater's SD data for each item were obtained from Hughes et al. (2021). The correlation between the scale parameter, a_j, and SD_j in this bank is .30 (with $r_{a_j^+ - SD_j^+} = .01$ and $r_{a_j^- - SD_j^-} = .21$). We assembled four questionnaires for each optimization method, two with a total of 70 blocks, where one had all homopolar blocks and one with 20% heteropolar blocks (14 blocks), and two with 35 blocks, where one had all homopolar blocks and one with 20% heteropolar blocks (7 blocks). All questionnaires incorporated SDM using a 0.75 cutoff for homopolar blocks and 1.125 for heteropolar blocks, since SD ratings ranged from 1 to 7. Likert item parameters were obtained with a graded response model. For the estimation of marginal reliability used in the BF and GA methods, we used the empirical NEO-PI-R factor correlation matrix (Costa & McCrae, 1992). For algorithm comparison, we considered empirical reliability estimates. The procedure was as follows. First, 5,000 responses to the FCQ were simulated using the a_j and d_j parameters derived from estimates based on the Likert-format version, along

with the estimated trait correlation matrix. Next, the MUPP-2PL model was fitted to the simulated data. Finally, empirical reliability was calculated using the empirical_rxx() function from the R package *mirt*.

As seen in Table 3, GA consistently provides the most reliable questionnaire scores (close to or above $\widehat{\rho}_{emp} = .70$), followed by SA with a_j optimization, which sometimes performs the same as BF, and SA blueprint. Differences were smaller when assembling longer questionnaires.

Table 3

Average Trait Recovery Using Each Algorithm in the Empirical Illustration

	$\widehat{\rho}_{emp}$			
Condition	**GA**	**SA$_{a_j}$**	**BF**	**SA$_{bp}$**
70-blocks				
20% heteropolar	**.78**	.77	.76	.74
0% heteropolar	**.79**	.76	.75	.74
35-blocks				
20% heteropolar	**.71**	.65	.64	.61
0% heteropolar	**.70**	.65	.63	.59

Note. Maximum values of $\widehat{\rho}_{emp}$ are marked in bold. GA = genetic algorithm; SA$_{a_j}$= simulated annealing with scale parameter, a_j, optimization; BF = brute-force; SA$_{bp}$ = simulated annealing blueprint.

Discussion

The optimal assembly of non-cognitive questionnaires with adequate reliability has become increasingly relevant with the growing use of FC formats. Despite its importance, no previous study has systematically compared the performance of existing methods for the optimal assembly of FCQs. Therefore, the present study addresses this gap by evaluating the impact of four assembly methods on trait score recovery through a simulation study and an empirical illustration: GA (Kreitchmann et al., 2022), with an improvement on the application of SD constraints; two approaches of SA implemented in the *autoFC* package (Li et al., 2022), one using a blueprint method and the other optimizing the within-block difference in the a_j parameters; and a BF random search. Two key aspects identified in the literature as critical for assembly are the use of heteropolar blocks (often recommended to avoid ipsativity issues) and the correlation between a_j and SD_j, which can make optimal pairing difficult (i.e., pairing items with different a_j, while keeping them matched in SD_j; Lee et al., 2022; Li et al., 2025; Pavlov et al., 2021). These factors, along with the questionnaire length, were incorporated into the simulation design and empirical illustration.

As we hypothesized, among the methods tested, GA (Kreitchmann et al., 2022) produced systematically the most reliable questionnaire scores in both the simulation study and empirical illustration. This result can be expected, as this method explicitly optimizes the reliability of the assembled blocks. However, closely behind is the SA with a_j parameter optimization which aligns with promising results found in Li et al. (2025). In terms of implementation, GA is more computationally demanding and slower, whereas SA is faster. The SA blueprint method performed similarly to or worse than the BF approach in the simulation study. This result is reasonable, as BF selects the questionnaire with the highest expected score reliability from the 100 generated, whereas the SA blueprint method does not include an explicit reliability optimization step. All examined factors significantly affected reliability, including questionnaire length, SDM, the degree of relation between a_j and SD_j, and the inclusion of heteropolar blocks. As expected, longer questionnaires yielded higher reliability, and the impact of the assembly algorithm became more pronounced with increased length. GA showed the greatest advantage under the most demanding conditions. The results for the ipsativity indicator, bias in the recovery of the correlation matrix, preserve the same ordering of methods as observed for reliability. Ipsativity bias tends to increase when heteropolar blocks are absent, but its magnitude depends on the scenario. In many conditions, including some with 0% heteropolar blocks, bias remains minimal across algorithms, and only under more demanding scenarios such as high correlations between scale parameters and social desirability, shorter tests, or less optimal assembly algorithms does it reach larger values. Across all conditions, GA consistently shows the lowest bias, followed by SA with a_j optimization, BF, and SA blueprint, highlighting that both test design and algorithm choice can influence the magnitude of distortion. This comparison of the algorithms in terms of this variable is informative, since none of them explicitly aims to minimize this bias, allowing for a fairer comparison.

In the simulation study, mixed keyed tests produced higher reliability than positively keyed tests alone. This further emphasizes that the incorporation of heteropolar blocks can be useful if paired correctly, supporting the findings of Brown and Maydeu-Olivares (2011) and Frick et al. (2023), but contrasts with those reported by Graña et al. (2025). This did not align with our second hypothesis. It should be noted that the criteria was more lenient when combining items with greater variability in SD in the case of heteropolar blocks, which may have contributed to the observed increase in reliability. Moreover, excluding heteropolar blocks reduces the search space. In the empirical example, this exclusion results in disregarding 12,392 item combinations. Reducing the search space also limits the flexibility to maximize reliability. If the applied researcher considers incorporating heteropolar blocks to be beneficial they can do so, as it can increase reliability, which can be achieved with only a few blocks. However, this can have a downside, since they are harder to form, especially when imposing other constraints such as SDM, as seen in the empirical illustration, when a smaller bank results in fewer

viable heteropolar blocks. Such restrictions are particularly problematic in high-stakes contexts where item leakage may further reduce usable item combinations.

In line with our third hypothesis, the relationship between item scale parameters and SD also proved to be relevant. In banks consisting only of positively keyed items, lower correlations between a_j and SD_j led to higher reliability. In contrast, high correlations (e.g., $r_{a_j^+ - SD_j^+} = .80$) reduced reliability; this is more noticeable when matching for SD, likely because highly discriminating items were also highly socially desirable, making it harder to assemble blocks with high a_j and minimal SD_j differences. These results indicate that the more difficult the SDM task is, the more the choice of method matters, with GA showing the best performance.

Regarding our fourth hypothesis, SDM tends to reduce reliability, which we hypothesized would happen because, as an additional constraint, it reduces reliability regardless of the algorithm. This effect is most pronounced in BF, whereas in GA and SA the loss in reliability is smaller. This decrease is typically less than .01, except for the most demanding condition ($r_{a_j^+ - SD_j^+} = .80$). Nevertheless, the difference remained small, and SDM remains important in practice, as it can be easily implemented and can enhance validity.

While the findings are promising, certain limitations must be considered and may inform future research efforts. A limitation of this study is its focus on pairwise blocks, leaving the evaluation of these assembly methods for larger block sizes to future research. This constraint also applies to GA, which currently does not support the formation of blocks containing more than two items. In this study, we rely on the assumption of measurement invariance, with supporting evidence provided by Morillo et al. (2019). Nevertheless, prior research acknowledges that this assumption may not hold universally (Kreitchmann et al., 2022) and may be influenced by block polarity (Graña et al., 2025). Accordingly, the reliance on measurement invariance represents a limitation of the current study. As any method that assembles blocks from SS item parameters inherently relies on this assumption, and because measurement invariance remains a critical issue, further empirical work is needed to delineate the cases where it may not be sustained. It may be valuable to explore alternative approaches such as psychometric networks to examine it in more depth (Abdelhamid et al., 2024; Jamison et al., 2024). Applied researchers interested in using the algorithms used in this study should consider this when assembling FCQs. Here, we focus on the MUPP-2PL model, although the proposed procedures can in principle be extended to the TIRT framework. The choice of model may have some impact, because even though MUPP-2PL and TIRT are nearly equivalent for binary FC data, they posit different response processes and may therefore yield slightly different block-level parameter estimates. Nevertheless, as our interest lies in the comparative performance of the block-assembly algorithms rather than in the absolute values of the parameters, the overall pattern of results is not expected to change, and the main conclusions should generalize to TIRT-based applications. Another potential

direction for future research is to extend the GA approach by incorporating alternative reliability metrics beyond marginal reliability as the objective function. For instance, GA could be adapted to prioritize the reliability of a specific trait or apply weighted importance to certain traits over others or ensure a minimum level of reliability for all traits, rather than simply maximizing the average reliability across all traits. Additionally, an interesting direction for future research would be the incorporation of reliability in the optimization of energy in the *autoFC* package. In this regard, a normative-order indicator was used here as a measure of performance (reliability and similarity between estimated and true theta), although other classification-related metrics may be of interest in applied settings. These other metrics can be investigated in future research. In line with this further exploration and with the goal of supporting practical use, the R functions used in this study have been made available in an OSF repository, at Sorrel et al. (2026). This will facilitate their application, especially for potential users of GA, who previously had to rely on the Shiny app. It would be interesting for future research to examine whether other possible options for block assembly, such as linear programming or ant colony algorithms, might offer better performance under certain conditions.

Conclusion

In conclusion, we recommend using the GA for assembling FCQs, as it consistently produces high-quality solutions by accounting for key psychometric properties. Its advantages are particularly evident in challenging scenarios such as short questionnaires, high correlation between a_j and SD_j, and the need to match items on SD. Although SDM may slightly reduce reliability due to its restrictive nature, it remains essential for minimizing response bias. As the forced-choice format continues to gain popularity over traditional SS formats, the use of optimization algorithms becomes increasingly important. These methods enable fast and reliable questionnaire assembly while accommodating additional constraints such as SD control and the inclusion of heteropolar blocks, making them an essential tool for advancing non-cognitive assessment.

Funding: This work was funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU under the project "Computerized adaptive tests based on new assessment formats" (Reference: PID2022-137258NB-I00) and the UAM-IIC Chair Psychometric Models and Applications.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors declare that there are no conflict of interests to disclose.

Author Contributions: *Scarlett Sophie Escudero*: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. *Miguel A. Sorrel*: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. *Rodrigo S. Kreitchmann*: Conceptualization, Methodology, Software, Validation, Writing – review & editing. *Francisco José Abad*: Conceptualization, Funding acquisition, Methodology, Resources, Validation, Writing – review & editing.

Data Availability: The data and code used in the manuscript are publicly available at Sorrel et al. (2026). Supplementary tables and figures to this study can be found at Escudero et al. (2026)

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
IPIP300.por	Sorrel et al. (2026)
Code	
Empirical illustration - R code	Sorrel et al. (2026)
Simulation functions - R code	Sorrel et al. (2026)
block Assembly NHBSA - R code	Sorrel et al. (2026)
NHBSA Functions - R code	Sorrel et al. (2026)
NHBSA - R code	Sorrel et al. (2026)
select Blocks - R code	Sorrel et al. (2026)
Material	
IPIP Neo ItemKey	Sorrel et al. (2026)
Study/Analysis preregistration	
Study was not preregistered	—
Other	
Algorithm Specifications	Sorrel et al. (2026)
Supplementary Tables and Figures	Escudero et al. (2026)

References

Abdelhamid, G. S. M., Hidalgo, M. D., French, B. F., & Gómez-Benito, J. (2024). Partitioning dichotomous items using Mokken scale analysis, exploratory graph analysis and parallel analysis: A Monte Carlo simulation. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 20*(3), 187–217. https://doi.org/10.5964/meth.12503

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

Bürkner, P.-C. (2022). On the information obtainable from comparative judgments. *Psychometrika, 87*(4), 1439–1472. https://doi.org/10.1007/s11336-022-09843-z

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

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. *Journal of General Psychology, 125*(3), 245–261. https://doi.org/10.1080/00221309809595548

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO-PI-R Professional Manual.* Psychological Assessment Resources.

Escudero, S., Sorrel, M. A., Kreitchmann, R. S., & Abad, F. J. (2026). *Supplementary Materials to* "A comparison of optimization algorithms for forced-choice questionnaire assembly" [Supplementary tables and figures to this study]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.22226

Frick, S., Brown, A., & Wetzel, E. (2023). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research, 58*(1), 1–29. https://doi.org/10.1080/00273171.2021.1938960

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg University Press.

Graña, D. F., Kreitchmann, R. S., Abad, F. J., & Sorrel, M. A. (2025). Equally vs. unequally keyed blocks in forced-choice questionnaires: Implications on validity and reliability. *Journal of Personality Assessment, 107*(3), 392–405. https://doi.org/10.1080/00223891.2024.2420869

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

Hughes, A. W., Dunlop, P. D., Holtrop, D., & Wee, S. (2021). Spotting the "ideal" personality response: Effects of item matching in forced choice measures for personnel selection. *Journal of Personnel Psychology, 20*(1), 17–26. https://doi.org/10.1027/1866-5888/a000267

Jamison, L., Christensen, A. P., & Golino, H. F. (2024). Metric invariance in exploratory graph analysis via permutation testing. *Methodology, 20*(2), 144–186. https://doi.org/10.5964/meth.12877

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Kreitchmann, R. S., Abad, F. J., & Sorrel, M. A. (2022). A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires. *Behavior Research Methods, 54*, 1476–1492. https://doi.org/10.3758/s13428-021-01677-4

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*, Article 2309. https://doi.org/10.3389/fpsyg.2019.02309

Lee, P., Joo, S. H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods. *Personality and Individual Differences, 191*, Article 111555. https://doi.org/10.1016/j.paid.2022.111555

Lee, P., Son, M., Zhou, S., Joo, S., Jia, Z., & Cheng, V. (2025). The journey of forced choice measurement over 80 years: Past, present, and future. *Organizational Research Methods, 28*(4), 680–722. https://doi.org/10.1177/10944281251350687

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. https://doi.org/10.1177/01466216211051726

Li, M., Zhang, B., Li, L., Sun, T., & Brown, A. (2025). Mixed-keying or desirability-matching in the construction of forced-choice measures? An empirical investigation and practical recommendations. *Organizational Research Methods, 28*(2), 296–329. https://doi.org/10.1177/10944281241229784

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/10.1177/0013164416646162

McCrae, R. R., Herbst, J. H., & Costa, P. T., Jr. (2001). Effects of acquiescence on personality factor structures. In R. Riemann, F. Ostendorf & F. Spinath (Eds.), *Personality and temperament: Genetics, evolution, and structure* (pp. 217–231). Pabst Science.

Morillo, D. (2018). *Item response theory models for forced-choice questionnaires* (Doctoral dissertation, Universidad Autónoma de Madrid). https://repositorio.uam.es/server/api/core/bitstreams/99f23341-5061-4586-a976-3a631ec7721d/content

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 multiunidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 40*(7), 500–516. https://doi.org/10.1177/0146621616662226

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

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

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

Sorrel, M. A., Escudero, S., Abad, F. J., & Kreitchmann, R. S. (2026). *A comparison of optimization algorithms for forced-choice questionnaire assembly* [OSF project page containing R code/functions used in the study, study data, document detailing all study algorithm specifications]. Open Science Framework. https://osf.io/2ubgh/overview

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