SoftwareX 36 (2026) 103157

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original Software Publication

OPTIVAL: An R package for model-based optimal test validity analysis

David Navarro-González *, Pere J. Ferrando, Fabia Morales-Vives

Department of Psychology, Research Center for Behavior Assessment (CRAMC), Universitat Rovira i Virgili, Tarragona, Spain

ARTICLE INFO

Keywords:
Optimal validity
Item selection
R package
Factor analysis

ABSTRACT

Maximizing external validity (the capacity of test scores to predict relevant criteria) is a central concern in psychological and educational measurement. This article presents OPTIVAL, an R package that implements a model-based approach for selecting item subsets that optimize the predictive validity of test scores with respect to an external criterion. Items are calibrated with a unidimensional factor-analytic model, model-implied validity curves and optimal item rankings are derived, and the accuracy of the predictions is checked through bootstrap resampling. The package provides both programmatic R functions and an interactive Shiny application, and is freely available on GitHub under a GPL-3 license. Its functioning is illustrated with simulated and real data examples.

Metadata

Nr	Code metadata description	Metadata
C1	Current code version	1.0.2
C2	Permanent link to code/repository used for this code version	https://github.com/dnavarrogonzalez/OPTIVAL/releases/tag/1.0.2
C3	Legal code license	GPL-3
C4	Code versioning system used	Git
C5	Software code languages, tools and services used	R, JavaScript, CSS, HTML
C6	Compilation requirements, operating environments and dependencies	R ($\geq$ 4.0.0); Dependencies: shiny ($\geq$ 1.7.0), lavaan ($\geq$ 0.6–0), DT ($\geq$ 0.20), plotly ($\geq$ 4.10.0), waiter ($\geq$ 0.2.0)
C7	If available, link to developer documentation/manual	https://github.com/dnavarrogonzalez/OPTIVAL/blob/main/README.md
C8	Support email for questions	david.navarro@urv.cat

1. Motivation and significance

Maximizing the external validity of test scores (their capacity to predict relevant criteria and other variables [1,2]) is a pervasive challenge in applied psychological and educational measurement. One of the most immediate determinants for improving predictive validity is test composition and length, a topic studied since at least the 1920s [3] through Classical Test Theory [4,5], covariance algebra [6,7], and purely empirical selection methods [8–10]. Empirical approaches, however, are known to capitalize on sampling error and calibration uncertainty [8–10], producing results that are unstable across samples and difficult to generalize across criteria.

A model-based approach, in which items are calibrated using a testable measurement model and item-criterion relations are next predicted from the resulting parameter estimates, addresses these limitations but has received comparatively little attention in applied measurement and, so far, has not produced self-contained software implementations. The theoretical framework underlying our proposal is developed in detail in Ferrando et al. [11,12]: the extended congeneric model incorporating the external criterion variable, the conditions under which additional items increase validity, and the model-based derivation of validity curves and optimal item subsets. The present article describes OPTIVAL, an R package that implements this proposal.

Several R packages already support automated item selection for short-form construction. ShortForm [13] implements ant colony optimization, tabu search and simulated annealing on top of lavaan; stuart [14] constructs subtests within a confirmatory factor-analytic framework by means of ant colony optimization, genetic algorithms or brute-force search; and GAabbreviate [15] uses genetic algorithms to find the item subset that best reproduces the full-scale scores. These are flexible, general-purpose optimizers: the user defines an objective function (e.g., model fit, reliability, or relations with external variables, which ShortForm and stuart can include) and the algorithm searches the space of candidate subsets empirically [16]. OPTIVAL differs in scope and logic. It addresses a single, well-defined target (the predictive validity of the unweighted sum of standardized item scores with respect to one external criterion) and, instead of searching the space of subsets, it

* Corresponding author.
E-mail address: david.navarro@urv.cat (D. Navarro-González).

https://doi.org/10.1016/j.softx.2026.103157
Received 23 July 2026; Received in revised form 6 October 2026; Accepted 15 October 2026
Available online 22 October 2026
2352-7110/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

derives the optimal order of entry, the expected validity curve and the stopping point analytically from the parameters of a testable measurement model. The selection is therefore deterministic given the calibration, does not require evaluating large numbers of candidate subsets on the same data, and comes with explicit model-based expectations (including a validity ceiling) that are checked against bootstrap evidence. Conversely, metaheuristic tools remain the better choice when several goals must be balanced simultaneously, such as content or facet coverage in multidimensional instruments.

In its current form, OPTIVAL is designed for the scenario described in Ferrando et al. [12]: a calibrated unidimensional measure, unweighted sums of standardized item scores, and a single continuous external criterion. Within this scope, the package automates item calibration via Confirmatory Factor Analysis (CFA), computes model-implied validity curves and optimal item rankings, validates predictions through nonparametric bootstrap resampling, and selects a subset that balances measurement efficiency with predictive adequacy. It is freely available on GitHub (https://github.com/dnavarrogonzalez/OPTIVAL) under a GPL-3 license and provides both direct R functions and a Shiny web application for users who prefer a graphical interface.

The intended users are applied researchers and practitioners (psychometricians, scale developers, and measurement-oriented scientists in psychology, education, and health sciences) who need to make model-based decisions about optimal test length in predictive contexts.

2. Software description

2.1. Software architecture

OPTIVAL follows standard R package structure with integrated Shiny application components (Fig. 1). The architecture is organized in three functional layers:

Core computational layer: This layer contains the core statistical functions. Key functions include bootsam2() for bootstrap resampling with replacement, compaboot() for computing bootstrap-based item-criterion validity coefficients, Burisch() for calculating model-based incremental validity, increvalsim() for empirical incremental validity from observed data, and simulboot() for bootstrap simulation of validity curves. These are internal (non-exported) functions: they are documented with roxygen2 and advanced users can inspect or call them through the ::: operator (e.g., `OPTIVAL:::Burisch()`), but the supported user-facing interface consists of the three exported functions: OPTIVAL(), run_optival(), which launches the Shiny application, and optival_example(), which returns the paths to the example datasets distributed with the package. In addition, the object returned by OPTIVAL() has a plot() method that draws the diagnostic plots described below.

Analysis layer: The main OPTIVAL() function serves as the primary user-facing interface, integrating all core computational functions into a complete analysis pipeline. This function manages the end-to-end workflow: preprocessing and validating input data, conducting CFA using the lavaan package [17] to obtain item structural estimates, computing both model-based and bootstrap-based validity indices for comparison, determining optimal item subsets using a hybrid selection criterion balancing efficiency and effectiveness, and packaging all results into a structured output object with comprehensive diagnostics including model fit statistics, item-level parameters, and validity curves.

User interface layer: For researchers who prefer graphical interaction over programming, OPTIVAL provides a Shiny web application that supports file upload of data matrices (CSV/TXT) and optional pre-calibrated factor loadings, configuration of bootstrap replications and data formatting, progress indicators, interactive plotly visualizations, downloadable results tables, and summary statistics with interpretive guidance. The application makes the complete analysis pipeline accessible to users without R programming experience. Fig. 2 shows the application after an analysis has been run.

2.2. Software functionalities

Theoretical background. OPTIVAL implements the model-based framework of Ferrando et al. [11,12], which we briefly summarize here. Let z_{ij} be the standardized score of respondent i on item j (j = 1, …, m), y_i the standardized score of respondent i on the external criterion, and θ_i the level of respondent i on the common factor measured by the items. The extended congeneric model [11,12,18,19] treats the items and the criterion as linear indicators of this same factor:

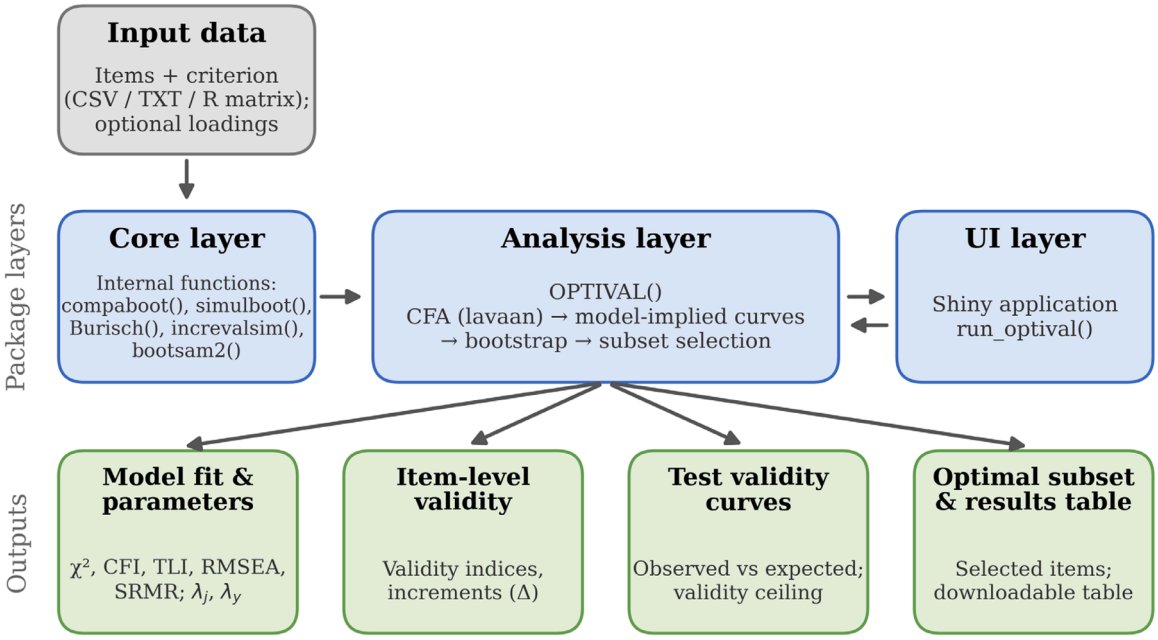


Fig. 1. OPTIVAL package architecture and analysis workflow. The package is organized in three functional layers (core computational, analysis, and user interface) that process input data and produce four types of output: model fit and parameter estimates, item-level validity results, test validity curves, and the optimal subset with a downloadable results table.

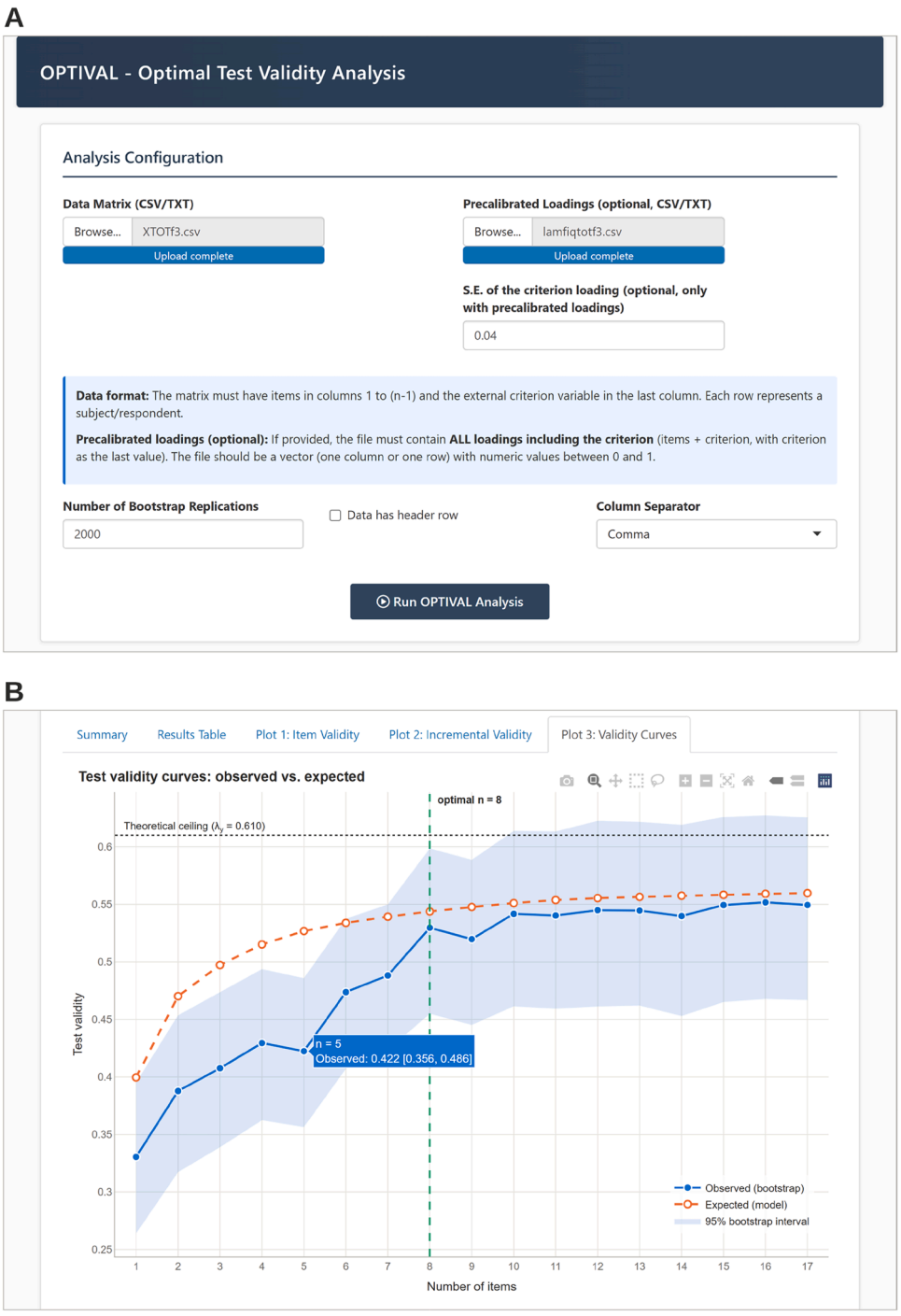


Fig. 2. The OPTIVAL Shiny application, launched with run_optival(), with the data of Example 2. (A) Analysis configuration: data matrix and precalibrated loadings uploaded, standard error of the criterion loading, and number of bootstrap replications. (B) Interactive validity curves after running the analysis; hovering over a point shows the observed validity at that test length with its 95% bootstrap interval. The other output tabs show the summary, the results table and the remaining two diagnostic plots.

$$z_{ij} = \lambda_j \theta_i + \varepsilon_{ij}, \quad y_i = \lambda_y \theta_i + \varepsilon_{iy} \tag{1}$$

where λ_j is the factor loading of item j, λ_y is the loading of the criterion (its structural validity coefficient), $\theta \sim N(0, 1)$, and the residuals ε_{ij} and ε_{iy} are mutually uncorrelated (local independence), with variances $1 - \lambda_j^2$ and $1 - \lambda_y^2$. Model (1) has two implications that OPTIVAL exploits. First, the correlation between item j and the criterion is

$$\rho_{jy} = \lambda_j \lambda_y \tag{2}$$

so that item validities are proportional to item loadings. Second, the validity of the unweighted sum of standardized item scores X_n of any subset of n items is

$$\rho(X_n, y) = \lambda_y \Sigma \lambda_j \Big/ \left[\left(\Sigma \lambda_j \right)^2 + \Sigma \left(1 - \lambda_j^2 \right) \right]^{1/2} = \lambda_y \omega_n^{1/2} \tag{3}$$

where the sums run over the n items in the subset and ω_n is the model-implied (omega) reliability of X_n. Eq. (3) shows that, for a given n, validity is maximized by the n items with the largest loadings, which defines the optimal order of entry, and that validity increases with n at a

decreasing rate towards λ_y, which is therefore the theoretical ceiling that no unweighted sum of these standardized items can exceed. To decide when to stop adding items, Ferrando et al. [12] re-express Burisch's covariance-algebraic condition for a validity increase [20] in terms of the model parameters: for the item entering in position $n+1$, the expected increment Δ_{n+1} is the amount by which its model-implied validity (Eq. 2) exceeds the minimum validity it would need for the validity of the sum score to increase. Positive values of Δ indicate a gain, and values close to zero indicate that further items are practically redundant. The functionalities described below implement these results.

OPTIVAL implements several interconnected functionalities that together provide a complete framework for optimal item selection:

- **Model-based calibration:** Items are calibrated using a unidimensional linear factor-analytic model, treated as a linear IRT model for continuous item responses [21]. The external criterion variable is then incorporated into an extended congeneric measurement model alongside the items [11,12,18,19]. From Eqs. (2) and (3), OPTIVAL derives model-implied item validity indices and predicts how test validity changes with the number of items included. The theoretical derivations are given in Ferrando et al. [11,12]; the software implements them directly.
- **Bootstrap validation:** To empirically verify model predictions and identify potential model misspecification, OPTIVAL employs nonparametric bootstrap resampling. The software draws repeated samples with replacement from the observed data, computing validity indices and test validity curves for each bootstrap replication. This provides empirical estimates of item-criterion relationships and their sampling variability, allowing comparison with model-based predictions. Discrepancies between model predictions and bootstrap estimates can reveal model misspecification or violations of assumptions. The number of bootstrap replications is user-configurable (default: 2000). Bootstrap standard errors and percentile confidence intervals are reported for the item validity indices and for the observed validity at each test length, and the confidence interval of the criterion loading λ_y is also reported, so that the uncertainty of the validity ceiling can be taken into account.
- **Incremental validity analysis:** OPTIVAL implements the incremental validity framework derived in Ferrando et al. [12], which embeds Burisch's covariance-algebraic result [20] within the measurement model. For each successive item added (ordered by loading), the expected increment Δ is computed and displayed graphically. The analysis identifies the point at which adding further items provides negligible or negative gains; this defines the stopping rule for item selection.
- **Optimal subset selection:** OPTIVAL employs a hybrid selection method that combines two complementary criteria. The efficiency criterion identifies items contributing less than a threshold proportion (default: 10%) of the maximum incremental validity, flagging items that provide minimal additional information. The effectiveness criterion ensures that the selected subset captures adequate total validity (default: 95% of the maximum of the bootstrap validity curve across all test lengths). The algorithm computes both criteria independently and selects the larger (more conservative) of the two cutoffs, so that both conditions are met simultaneously. Both thresholds are arguments of OPTIVAL() (efficiency_threshold and effectiveness_threshold) and can be changed by the user. Because items always enter in the model-implied order, the thresholds determine only the length of the selected subset, not which items are prioritized, so the subsets obtained under different thresholds are nested. A sensitivity analysis of the default values is reported in Section 3.3.
- **Comprehensive diagnostics:** OPTIVAL reports complete structural fit statistics (χ^2, CFI, TLI, RMSEA, SRMR), item-level results (factor loadings, bootstrap-estimated validity indices with sampling variability, incremental contributions, and selection order), and test-level results (the optimal subset, observed and model-expected validity curves, the correlation between loadings and validity indices, the theoretical validity ceiling, and efficiency metrics relative to this ceiling). These outputs allow users to assess both the appropriateness of the model and the quality of the resulting selection.

All results are returned by OPTIVAL () as a list of class "OPTIVAL", whose components are organized in five blocks. (i) Data information: `n_items`, `n_subjects` and `n_bootstrap`. (ii) Calibration results: `item_loadings` (λ_j), `criterion_loading` (λ_y, the validity ceiling) and `criterion_loading_ci` (its standard error and confidence interval, taken from the lavaan solution or, with precalibrated loadings, computed from a standard error supplied through the argument `criterion_loading_se`), `item_order` (the items sorted by decreasing loading, that is, the order of entry), and `model_fit` (χ^2, df, p, CFI, TLI, RMSEA with its 90% confidence interval, and SRMR; NULL when precalibrated loadings are supplied). (iii) Item-level validity results: `validity_indices` (bootstrap mean item-criterion correlations) with their standard errors and percentile confidence intervals (`validity_indices_se`, `validity_indices_ci`), validity_correlation (the correlation between loadings and validity indices), `validity_lm` (the corresponding linear fit), and `incremental_validity` (Δ for each position in the order of entry). (iv) The selected subset: `optimal_n_items`, `optimal_items`, and selection_diagnostics, a list with the lengths suggested separately by the efficiency and effectiveness criteria, the criterion that determined the final length, the thresholds used, and the validity reached at each suggested length. (v) Test-level validity curves: `observed_validity` (with `observed_validity_se` and `observed_validity_ci`) and `expected_validity` for every test length, `max_observed_validity` and `max_observed_n`, and `validity_by_n`, a data frame gathering the observed values and their confidence limits, the expected values and the increments by test length. The confidence level of all intervals is set with the argument `ci_level` (default 0.95). The plot() method draws the three diagnostic plots from this object (all together or one at a time, through its argument which), and the Shiny application displays interactive versions of the same plots, with the same visual encoding, together with a downloadable table.

2.3. Implementation details

OPTIVAL is implemented entirely in R, with the Shiny interface components using JavaScript for client-side interactivity and CSS for improved styling. The package depends on several well-established R packages available through the Comprehensive R Archive Network (CRAN): lavaan for structural equation modeling and confirmatory factor analysis [17], shiny for the web application framework, plotly for interactive data visualizations with zoom and hover capabilities, DT for enhanced, searchable data tables, and waiter for loading animations during computation.

The computational core relies on vectorized R operations and R's optimized linear algebra routines (BLAS/LAPACK). For typical applications ($n = 200$–1000 respondents, 10–50 items), analyses complete within seconds to minutes on a standard desktop computer; very large samples or bootstrap counts may extend computation to several minutes.

The package adheres to standard R packaging conventions. All exported functions include roxygen2 documentation with parameter descriptions, return value documentation, and usage examples. A README file provides installation instructions and quick-start examples. OPTIVAL is released under GPL-3, which ensures it remains freely available and permits modifications by the user community.

3. Illustrative examples

We present two illustrative examples: the first uses simulated data

with known structure to verify that the software recovers the expected relationships; the second applies OPTIVAL to a real dataset from educational psychology.

3.1. Example 1: simulated data with known structure

This example demonstrates the basic workflow using simulated data in which the true model structure is known.

Data generation: We simulated responses from 500 subjects to 20 items plus one external criterion. Items had factor loadings ranging from $\lambda = 0.25$ to $\lambda = 0.60$, representing a realistic spread of item quality for a noncognitive measure. The criterion had a population loading of $\lambda_y = 0.55$. Item responses were generated using the linear factor model $z_{ij} = \lambda_j\theta_i + \varepsilon_{ij}$ where $\theta_i \sim N(0,1)$ and $\varepsilon_{ij} \sim N(0, 1-\lambda_j^2)$.

The R code implementing this simulation and analysis is:

```
set.seed(123)
n <- 500; k <- 20
loadings <- seq(0.60, 0.25, length.out = k)
theta <- rnorm(n)
items <- sapply(loadings, function(lambda) {
  lambda * theta + sqrt(1 - lambda^2) * rnorm(n)
})
criterion <- 0.55 * theta + sqrt(1 - 0.55^2) * rnorm(n)
data_matrix <- cbind(items, criterion)
results <- OPTIVAL(data_matrix, n_bootstrap = 2000)
```

Model-data fit was good, as it should be: $\chi^2(189) = 205.14$, $p = 0.20$; CFI = 0.988, TLI = 0.987; RMSEA = 0.013 [0.000, 0.024]; SRMR = 0.033. The estimated criterion loading was $\lambda_y = 0.515$ (95% CI [0.442, 0.588]), representing the theoretical validity ceiling. OPTIVAL identified 10 items out of 20 as the optimal subset (items 2, 1, 6, 3, 4, 8, 7, 5, 10, 11 in order of incremental contribution). The bootstrap validity for the optimal subset was r = 0.454 (95% bootstrap CI [0.388, 0.520]), closely matching the model-expected value of r = 0.456. The full 20-item scale reached r = 0.466 [0.399, 0.533], so the 10-item version retains 97.5% of that value and operates at 88.2% efficiency relative to the theoretical ceiling.

Fig. 3 shows the three diagnostic outputs produced by OPTIVAL for this example. The figure was obtained directly with plot(results); the Shiny application displays interactive versions of the same three plots. Panel A plots item factor loadings against bootstrap-estimated item validity indices, with a linear trend (r = 0.827) consistent with the model prediction that loadings and validity indices should be proportional. The scatter around the trend reflects bootstrap variability, which is that expected in samples of this size. Panel B displays the expected incremental validity per item in order of entry: the first item contributes 0.309, the second 0.093, and contributions become negligible after the 10th, which is where the optimal cutoff is placed. Panel C shows the observed (bootstrap) and model-expected validity curves across all test

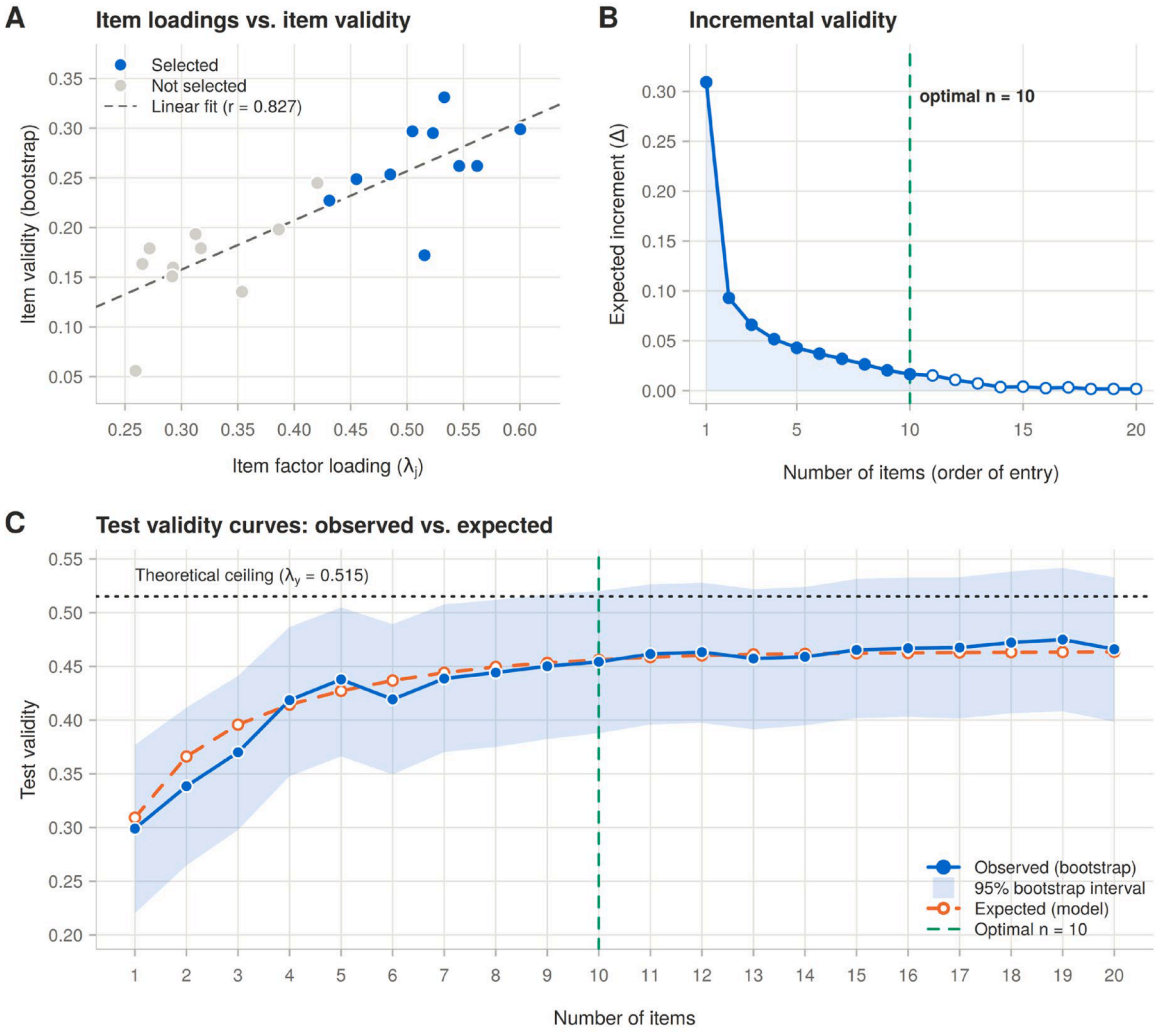


Fig. 3. Example 1 (simulated data, N = 500, k = 20). Panel A: Item factor loadings vs. bootstrap-estimated item validity indices (r = 0.827); filled circles denote the selected items. Panel B: Expected incremental validity per item in order of entry; optimal cutoff at n = 10. Panel C: Observed (bootstrap) and model-expected validity curves; the shaded band is the 95% bootstrap percentile interval of the observed validity; dotted line marks the theoretical ceiling ($\lambda_y = 0.515$).

lengths, together with the theoretical ceiling ($\lambda_y = 0.515$). The two curves match each other reasonably well, and both flatten visibly from n = 10 onwards.

3.2. Example 2: predicting family adaptability from home-based parental involvement

This real-data example uses a parental involvement measure as predictor of a family functioning criterion, and illustrates the use of precalibrated loadings from a previous analysis.

Dataset and context: The data are from Dueñas et al. (2022) [22] and comprise responses from 928 Spanish parents of high-school students to the Home-based activities subscale of the Family Involvement Questionnaire-High School (FIQ-HS). The external criterion is the Adaptability subscale of the Family Adaptability and Cohesion Evaluation Scales (FACES; Olson et al., 1982 [23]). The Home-based activities subscale (17 items, 4-point Likert format) assesses parental engagement in out-of-school activities that support learning, such as conversations about career options or help with homework. The FACES Adaptability subscale (10 items) measures a family system's ability to change its roles and rules in response to situational stress, and has been previously reported to be positively associated with parental involvement [22].

Item calibration: Rather than re-running CFA internally, OPTIVAL was provided with precalibrated factor loadings obtained from a Mplus analysis of the extended congeneric model using robust (MLMV) estimation, following the procedure described in Ferrando et al. [12]. This is the recommended approach when a high-quality external calibration is already available, and it avoids the fit degradation that ML estimation can produce with Likert-format items in large samples. Factor loadings ranged from $\lambda = 0.34$ to $\lambda = 0.66$. The structural validity coefficient for FACES Adaptability was $\lambda_y = 0.61$ (95% CI [0.53, 0.69], based on the Mplus standard error of 0.040). Full model fit statistics for this solution are reported in Ferrando et al. [12].

The data and loadings are distributed with the package, and the analysis below can be reproduced with:

```
d2 <- as.matrix(read.csv(optival_example("example2"), header = FALSE))
l2 <- read.csv(optival_example("example2_loadings"), header = FALSE)[, 1]
set.seed(123)
results2 <- OPTIVAL(d2, precalibrated_loadings = l2, n_bootstrap = 2000,
criterion_loading_se = 0.040)
```

OPTIVAL analysis and results: Running OPTIVAL with 2000 bootstrap replications and precalibrated loadings identified 8 items as the optimal subset (items 17, 15, 5, 4, 11, 7, 14, 8 in order of incremental

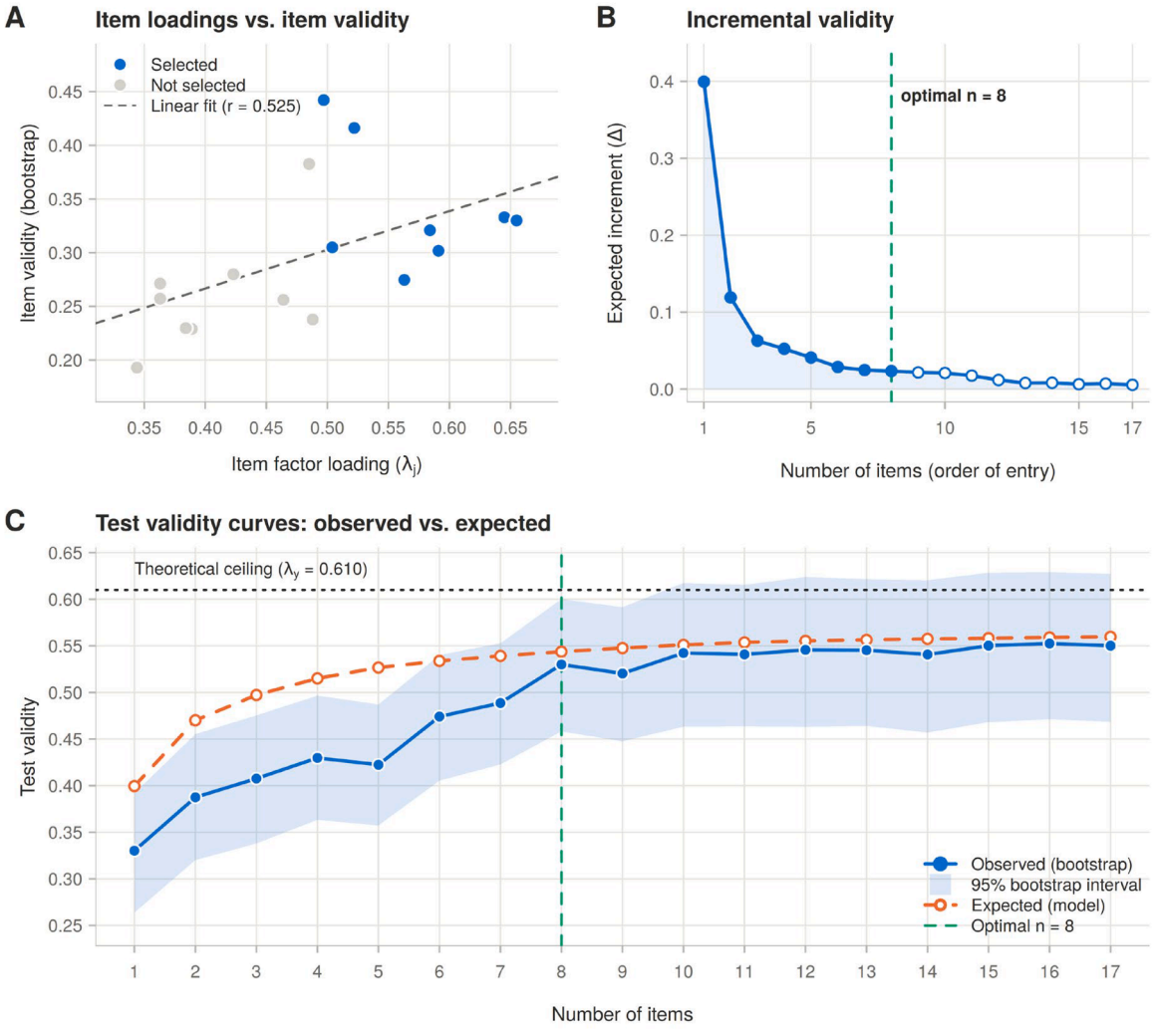


Fig. 4. Example 2 (FIQ-HS Home-based activities subscale, N = 928, k = 17; precalibrated loadings). Panel A: Item factor loadings vs. bootstrap-estimated item validity indices (r = 0.525); filled circles denote the selected items. Panel B: Expected incremental validity per item in order of entry; optimal cutoff at n = 8. Panel C: Observed (bootstrap) and model-expected validity curves; the shaded band is the 95% bootstrap percentile interval of the observed validity; dotted line marks the theoretical ceiling ($\lambda_y = 0.61$).

contribution). The bootstrap validity for this subset was r = 0.530 (95% bootstrap CI [0.458, 0.600]), against a model-expected value of r = 0.544 and a full-scale bootstrap validity of r = 0.550 [0.469, 0.627]. The 8-item version retains 96.3% of the full-scale validity while reducing test length by 53%, and operates at 86.9% efficiency relative to the theoretical ceiling of $\lambda_y = 0.61$. All selected items had loadings above $\lambda = 0.49$, well above the scale median.

Model-based insights: Three aspects of the output are worth highlighting. First, the correlation between item loadings and bootstrap-estimated item validity indices was r = 0.525 (Panel A of Fig. 4). This is lower than in the simulated example, as expected with real data: item validity indices reflect additional sources of variability beyond the common factor, so the proportionality predicted by the model only holds approximately. Second, the theoretical ceiling $\lambda_y = 0.61$ is not reached by the observed bootstrap curve, which levels off around r = 0.55; this reflects the unavoidable attenuation due to measurement error in the sum score. Third, the model-expected and bootstrap-observed validity curves converge from n = 6 onwards, where the model-expected values fall within the 95% bootstrap interval (Panel C), supporting the appropriateness of the pre-calibrated solution for these data. For the first five items, however, the model-expected values lie above the bootstrap interval: the items with the highest loadings (particularly items 17 and 15) are somewhat less predictive of the criterion than the model implies. This is precisely the kind of discrepancy that the bootstrap layer is meant to reveal; in this case it does not affect the selected subset, which lies in the region where both curves agree.

Practical implications: A 17-item questionnaire can be reduced to 8 items (53% shorter) while retaining 96.3% of its predictive power with respect to family adaptability: a practically relevant reduction in multi-battery surveys. The validity curve flattens quickly: most of the achievable validity is captured by the first four or five items, and items beyond eight add negligible information. The diagnostic output makes this tradeoff explicit and reproducible (Tables 1 and 2).

3.3. Sensitivity to the selection thresholds

Because the default thresholds (10% for efficiency and 95% for effectiveness) determine the length of the selected subset, we examined how the solution changes under alternative values. Table 3 reports the selected length for the nine combinations of efficiency (5%, 10%, 15%) and effectiveness (90%, 95%, 99%) thresholds, computed from the same OPTIVAL runs reported above. As noted in Section 2.2, changing the thresholds only moves the stopping point along the same model-implied ranking, so all the resulting subsets are nested. With effectiveness thresholds of 90% or 95%, the selected lengths ranged from 5 to 11 items in Example 1 and from 8 to 11 items in Example 2, and all of them retained at least 94% (Example 1) and 96% (Example 2) of the full-scale validity; in Example 2, the default length was obtained under four of these six combinations. The most influential setting was the 99% effectiveness threshold, which led to 18 and 15 items, respectively. Near the plateau of the validity curve, such a demanding target is only met once the small fluctuations of the bootstrap curve are overcome, so the additional items have small expected increments ($\Delta \leq 0.015$ in Example 1 and $\Delta \leq 0.022$ in Example 2) and add little validity. Finally, the selected length at the default thresholds was identical in 30 independent repetitions of each analysis with different random seeds (n = 10 in Example 1 and n = 8 in Example 2), so the bootstrap component does not make the selection unstable. Overall, the defaults provide a reasonable compromise between brevity and validity; when the decision is consequential, we recommend reporting the thresholds used together with a similar check, which only requires changing the two arguments of OPTIVAL().

Table 1
Item-level results for Example 1 (simulated data, N = 500, k = 20). Items in order based on increment in validity. Selected items (n = 10) are those retained in the optimal subset; validity ceiling $\lambda_y = 0.515$ (95% CI [0.442, 0.588]). Bootstrap item validity indices are shown with their 95% percentile intervals.

Rank	Item	Loading (λ_j)	Validity (Bootstrap) [95% CI]	Incr. Validity (Δ)	Selected
1	2	0.6004	0.299 [0.216, 0.374]	0.3092	✓
2	1	0.5621	0.262 [0.182, 0.333]	0.0930	✓
3	6	0.5462	0.262 [0.182, 0.338]	0.0660	✓
4	3	0.5330	0.331 [0.254, 0.404]	0.0517	✓
5	4	0.5229	0.295 [0.204, 0.379]	0.0429	✓
6	8	0.5155	0.172 [0.092, 0.250]	0.0371	✓
7	7	0.5047	0.297 [0.212, 0.378]	0.0319	✓
8	5	0.4851	0.253 [0.178, 0.328]	0.0264	✓
9	10	0.4551	0.249 [0.158, 0.332]	0.0204	✓
10	11	0.4312	0.227 [0.149, 0.309]	0.0166	✓
11	9	0.4206	0.245 [0.155, 0.324]	0.0152	–
12	13	0.3863	0.198 [0.116, 0.281]	0.0108	–
13	12	0.3537	0.135 [0.055, 0.219]	0.0073	–
14	17	0.3173	0.179 [0.102, 0.255]	0.0036	–
15	18	0.3127	0.193 [0.108, 0.271]	0.0041	–
16	15	0.2927	0.160 [0.077, 0.240]	0.0026	–
17	19	0.2918	0.151 [0.071, 0.228]	0.0033	–
18	20	0.2718	0.179 [0.097, 0.259]	0.0018	–
19	14	0.2655	0.163 [0.073, 0.252]	0.0018	–
20	16	0.2593	0.056 [−0.032, 0.141]	0.0018	–

Table 2
Item-level results for Example 2 (FIQ-HS Home-based activities subscale, N = 928, k = 17; precalibrated loadings). Items in order based on increment in validity. Selected items (n = 8) are those in the optimal subset; validity ceiling $\lambda_y = 0.61$, criterion = FACES Adaptability. Bootstrap item validity indices are shown with their 95% percentile intervals.

Rank	Item	Loading (λ_j)	Validity (Bootstrap) [95% CI]	Incr. Validity (Δ)	Selected
1	17	0.655	0.330 [0.262, 0.397]	0.3996	✓
2	15	0.645	0.333 [0.264, 0.403]	0.1191	✓
3	5	0.591	0.302 [0.232, 0.368]	0.0628	✓
4	4	0.584	0.321 [0.258, 0.379]	0.0524	✓
5	11	0.563	0.275 [0.210, 0.340]	0.0410	✓
6	7	0.522	0.416 [0.349, 0.484]	0.0288	✓
7	14	0.504	0.305 [0.241, 0.367]	0.0248	✓
8	8	0.497	0.442 [0.329, 0.539]	0.0234	✓
9	16	0.488	0.238 [0.172, 0.304]	0.0217	–
10	6	0.485	0.383 [0.262, 0.492]	0.0210	–
11	9	0.464	0.256 [0.187, 0.322]	0.0175	–
12	10	0.423	0.280 [0.164, 0.389]	0.0118	–
13	2	0.389	0.229 [0.161, 0.292]	0.0079	–
14	3	0.384	0.230 [0.134, 0.328]	0.0082	–
15	12	0.363	0.271 [0.210, 0.333]	0.0063	–
16	13	0.363	0.257 [0.192, 0.316]	0.0071	–
17	1	0.344	0.193 [0.126, 0.257]	0.0055	–

Table 3
Selected test length under alternative selection thresholds (in parentheses, bootstrap validity of the selected subset as a percentage of the full-scale bootstrap validity; values above 100% occur because the bootstrap validity curve is not strictly monotone). † Default thresholds.

Efficiency threshold	Effectiveness threshold	Example 1: n (% of full-scale validity)	Example 2: n (% of full-scale validity)
5%	90%	11 (99.0)	11 (98.3)
10%	90%	8 (95.3)	8 (96.3)
15%	90%	5 (94.0)	8 (96.3)
5%	95%	11 (99.0)	11 (98.3)
10%	95%	10 (97.5) †	8 (96.3) †
15%	95%	10 (97.5)	8 (96.3)
5%	99%	18 (101.3)	15 (100.0)
10%	99%	18 (101.3)	15 (100.0)
15%	99%	18 (101.3)	15 (100.0)

4. Impact

OPTIVAL addresses a gap in available psychometric tools. The core theoretical results underlying the package (validity-optimal composite construction within the linear factor model [19] and related external validity procedures [11,24]) have been available in the literature for some time, but translating them into practice requires coordinating item calibration, model-based prediction, and bootstrap validation in a single coherent pipeline. This is what OPTIVAL provides.

The most direct application is scale development and shortening: given a larger item pool, OPTIVAL identifies a subset that retains adequate predictive validity while reducing respondent burden. Unlike purely empirical selection on observed item-criterion correlations, which capitalizes on sampling error and produces unstable results across samples [8,25], OPTIVAL's predictions are derived from factor loadings, which, provided that the model is appropriate, tends to produce more stable and generalizable selections [19,24].

In validation studies, OPTIVAL provides a principled basis for justifying instrument length: rather than selecting the items with the highest item-total correlations, researchers can report a model-based analysis, with explicit efficiency and effectiveness thresholds, and supported by bootstrap evidence. This is relevant for test construction in personality and educational measurement, and even more so in clinical and diagnostic contexts [26].

The dual interface was a deliberate design choice: the Shiny application makes the pipeline accessible to researchers without programming experience, while the OPTIVAL() function integrates into programmatic workflows alongside packages like lavaan [17]. Both interfaces produce the same output.

Planned extensions include support for multidimensional structures, ordinal response formats, and multiple simultaneous criteria.

5. Conclusions

OPTIVAL implements model-based optimal validity analysis for noncognitive measures, grounding item selection in extended factor-analytic calibration rather than purely empirical correlations, and brings together calibration, model-implied validity curves, bootstrap validation, and subset selection in a single workflow.

The core contribution is an integrated implementation of the theoretical framework developed in Ferrando et al. [11,12], with the addition of a hybrid selection criterion and a Shiny application that broadens access to researchers without programming backgrounds. The bootstrap validation layer provides empirical checks on model predictions and makes visible any discrepancy between theoretical and observed validity curves.

The current version is limited to unidimensional structures and a single external criterion. Extending the framework to multidimensional models, ordinal response formats, and multiple simultaneous criteria are the most natural next steps; the modular architecture of the package is intended to support these additions.

A further limitation, inherent to the method, is that items are ranked only by their loadings and, therefore, by their expected contribution to the prediction of the criterion; item content plays no role in the selection. When a scale was designed to sample several facets of a construct, facet-specific items do not necessarily have the highest loadings, so the optimal subset may under-represent, or even exclude, some facets: a gain in predictive accuracy (fidelity) may come at the cost of content coverage (bandwidth). This trade-off is shared by any selection procedure driven by factor loadings or item-criterion correlations, and OPTIVAL is not meant to replace content-based judgment. We therefore recommend inspecting the content of the selected items and, when content coverage is essential, adjusting the subset accordingly or using procedures that allow explicit content constraints, such as those discussed in Section 1.

By reducing the technical barriers to model-based validity analysis, OPTIVAL aims to encourage more well-grounded approaches to test construction and shortening in psychology, education, and related fields where measurement quality matters for downstream decisions.

Data availability

The datasets used in the illustrative examples are distributed with the package (inst/extdata) and are available in the GitHub repository. The real data come from Dueñas et al. (2022) [22]. An R script that reproduces all the analyses, tables and figures reported in this article is also available in the repository.

CRediT authorship contribution statement

David Navarro-González: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Pere J. Ferrando:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Fabia Morales-Vives:** Writing – review & editing, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Grucza RA, Goldberg LR. The comparative validity of 11 modern personality inventories: predictions of behavioral acts, informant reports, and clinical indicators. J Pers Assess 2007;89(2):167–87. https://doi.org/10.1080/00223890701468568.
[2] Soto CJ, John OP. Optimizing the length, width, and balance of a personality scale: how do internal characteristics affect external validity? Psychol Assess 2019;31(4):444–59. https://doi.org/10.1037/pas0000586.
[3] Edgerton HA, Toops HA. A table for predicting the validity and reliability coefficients of a test when lengthened. J Educ Res 1928;18(3):225–34. https://doi.org/10.1080/00220671.1928.10879881.
[4] Gulliksen H. Theory of mental tests. New York: John Wiley & Sons; 1950.
[5] Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
[6] Gleser GC, DuBois PH. A successive approximation method of maximizing test validity. Psychometrika 1951;16(1):129–39. https://doi.org/10.1007/BF02313432.
[7] Penev S, Raykov T. Maximal reliability and power in covariance structure models. Br J Math Stat Psychol 2006;59(1):75–87. https://doi.org/10.1348/000711005X68183.
[8] Reiter-Palmon R, Connelly MS. Item selection counts: a comparison of empirical key and rational scale validities in theory-based and non-theory-based item pools. J Appl Psychol 2000;85(1):143–51. https://doi.org/10.1037/0021-9010.85.1.143.
[9] Bell RC, Lumsden J. Test length and validity. Appl Psychol Meas 1980;4(2):165–70. https://doi.org/10.1177/014662168000400203.
[10] Burisch M. Test length and validity revisited. Eur J Pers 1997;11(4):303–15. https://doi.org/10.1002/(SICI)1099-0984(199711)11:4<303::AID-PER292>3.0.CO;2-#.
[11] Ferrando PJ, Morales-Vives F, Durán-Bonavila S, Navarro-González D. Assessing the unconditional and conditional external validity of noncognitive test scores: a unifying model-based proposal. Educ Psychol Meas 2026. https://doi.org/10.1177/00131644261440168.
[12] P.J. Ferrando, D. Navarro-González, F. Morales-Vives, Revisiting the test-length and validity problem from a model-based view: a proposal and recommendations, An Psicol 2026.
[13] Raborn AW, Leite WL. ShortForm: an R package to select scale short forms with the ant colony optimization algorithm. Appl Psychol Meas 2018;42(6):516–7. https://doi.org/10.1177/0146621617752993.
[14] M. Schultze, stuart: subtests using algorithmic rummaging techniques, R package version 0.10.2, 2023. https://doi.org/10.32614/CRAN.package.stuart.
[15] Sahdra BK, Ciarrochi J, Parker P, Scrucca L. Using genetic algorithms in a large nationally representative American sample to abbreviate the multidimensional experiential avoidance questionnaire. Front Psychol 2016;7:189. https://doi.org/10.3389/fpsyg.2016.00189.
[16] Schroeders U, Wilhelm O, Olaru G. Meta-heuristics in short scale construction: ant colony optimization and genetic algorithm. PLoS ONE 2016;11(11):e0167110. https://doi.org/10.1371/journal.pone.0167110.
[17] Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Softw 2012;48(2):1–36. https://doi.org/10.18637/jss.v048.i02.

[18] Rosenbaum PR. Criterion-related construct validity. Psychometrika 1989;54(4):625–33. https://doi.org/10.1007/BF02296400.
[19] Ferrando PJ. Maximizing the information and validity of a linear composite in the factor analysis model for continuous item responses. Psicológica 2008;29(2):189–203.
[20] Burisch M. You don't always get what you pay for: measuring depression with short and simple versus long and sophisticated scales. J Res Pers 1984;18(1):81–98. https://doi.org/10.1016/0092-6566(84)90040-0.
[21] Ferrando PJ. Difficulty, discrimination, and information indices in the linear factor analysis model for continuous item responses. Appl Psychol Meas 2009;33(1):9–24. https://doi.org/10.1177/0146621608314608.
[22] Dueñas JM, Morales-Vives F, Camarero-Figuerola M, Tierno-García JM. Spanish adaptation of the family involvement questionnaire-high school: version for parents. Psicol Educ 2022;28(1):31–8. https://doi.org/10.5093/psed2020a21.
[23] Olson DH, Portner J, Bell R. FACES II: family adaptability and cohesion evaluation scales. Fam Soc Sci 1982.
[24] Ferrando PJ, Lorenzo-Seva U. An external validity approach for assessing essential unidimensionality in correlated-factor models. Educ Psychol Meas 2019;79(3):437–61. https://doi.org/10.1177/0013164418824755.
[25] van der Linden WJ, Glas CAW. Capitalization on item calibration error in adaptive testing. Appl Meas Educ 2000;13(1):35–53. https://doi.org/10.1207/S15324818AME1301_2.
[26] Borsboom D. Psychometric perspectives on diagnostic systems. J Clin Psychol 2008;64(9):1089–108. https://doi.org/10.1002/jclp.20503.