It's how you use the items that counts: An intelligent procedure for item selection in Item Response Theory

Ottavia M. Epifania[1,2], Pasquale Anselmi[2], & Egidio Robusto[2]

[1]Department of Psychology and Cognitive Sciences, University of Trento
[2]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova

Abstract

Item Response Theory (IRT) provides the ideal framework for generating short test forms (STFs) from item banks or full-length tests. The usual procedure for developing STFs is based on the visual inspection of the item and test information functions (IIFs and TIFs) to find the items that best contribute to make up a STF with the desired characteristics. This contribution presents a new procedure that aims at automating this process by defining a target TIF and finding the items from the item bank that best recover it. The algorithm directly compares the distance between the target TIF and a temporary TIF obtained by adding an item at a time. The items are chosen according to their closeness to the location on the latent trait where the distance between the target and temporary TIFs is maximum. The algorithm stops when the addition of a new item does not reduce the distance between target and temporary TIFs. The procedure can be applied both when the length of the STF is defined a priori and the optimal number of items has to be find. The results of the application of this procedure are presented.

Keywords: Item Response Theory, Test Information Function, short test forms, intelligent search algorithm

Introduction

Item Response Theory (IRT) models allow for estimating the probability that each person will endorse an item given their latent trait level and the characteristics of the items, as described by different parameters. The details at the item level inform about the precision with which each item measures different levels of the latent trait. As such, IRT models are of great use for developing short test forms (STFs), given the possibility they provide of minimizing the number of administered items while maximizing measurement precision.

Different procedures for the development of STFs based on IRT models exist, which are aimed at either the adaptive selection of items (i.e., each person is administered a different subset of items tailored to their specific latent trait level) or the fixed selection of items (i.e., static STFs where all people are administered the same subset of items). This contribution focuses on the development of static STFs. As such, if not otherwise specified, STFs identifies static STFs. Recently, Epifania, Anselmi, and Robusto (2022) introduced a new procedure for the development of STFs. This procedure is based on the a priori definition of the so-called θ targets, which are the levels of the latent trait on which the measurement precision should be focused the most. Although the procedure proved its soundness in developing STFs able to precisely measure the entire latent trait, the number of items to include in the STFs must be known in advance (it actually equals the number of θ targets).

In this contribution, a new procedure is introduced, which aims at developing STFs with characteristics that match those of a theoretically defined information function describing the characteristics of the desired test. As such, the number of items does not need to be set in advance. The procedure is validated through simulation studies and its performance is compared against a "brute force" procedure. The manuscript is organized as follows: the following section presents the 2-Parameter logistic model (2-PL, Birnbaum, 1968) and the information functions of the items and of the test. The descriptions of the newly introduced procedure and of the brute force procedure follow. Then, the simulation study is presented along with its results. A brief section illustrating the potentials and pitfalls of this procedure concludes the argumentation.

Item and Test Information Function in Item Response Theory

Different IRT models are available according to the nature of the observed responses (i.e., dichotomous vs. polytomous) and to the level of granularity desired for describing the functioning of the items (i.e., the number of item parameters). In this application, we refer to the 2-PL model (Birnbaum, 1968) for dichotomous responses. According to the 2-PL model, the probability of observing a correct response on item i ($i = 1, 2, \dots N$) provided by person p ($p = 1, 2, \dots, P$) is

$$P(x_{pi} = 1|\theta_p, b_i, a_i) = \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]},$$

where θ_p describes the latent trait level of person p (also known as ability parameter), b_i is the difficulty of item i (i.e., the location of the item on the latent trait), and a_i is the discrimination parameter of item i, which describes the ability of the item to discriminate between respondents with different θ levels. The probability of a correct response depends on the distance on the latent trait between the person parameter θ_p and the item location b_i, such that if $\theta_p > b_i$ this probability is greater than $.50$, while if $\theta_p < b_i$ the same probability is lower than $.50$. This probability is further influenced by the discrimination parameter, such that the higher the value of a_i, the more two respondents with similar θ levels have different probabilities of endorsing the item.

The *item information function* (*IIF*), $IIF_i = a_i^2[P(\theta, b_i, a_i)(1 - P(\theta, b_i, a_i))]$, expresses the precision with which each item measures each level of the latent trait θ. In the 2-PL model, the *IIF* is strongly influenced by the discrimination parameter (i.e., the higher the value a_i of item i, the higher its IIF) and it reaches its maximum (i.e., the item is mostly precise) when the ability θ_p matches the location b_i of item i, such that $P(\theta, b_i, a_i) = 1 - P(\theta, b_i, a_i)$.

By summing up the IIF of all the items, the *test information function* (*TIF*), $TIF = \sum_{i=1}^{N} IIF_i$, is obtained, which expresses the measurement precision of the test as a whole. Being the sum of IIFs, the TIF strongly depends on the number of items included in a test, and it can be increased just by adding new items. Given that the aim of the procedure introduced in this study is to minimize the number of administered items while recovering as best as possible the characteristics of a specific test, the use of the TIF as a criterion for choosing the most informative STF is misleading and would risk to favor STFs with a higher number of items. As such, the mean TIF, $\overline{TIF} = \sum_{i=1}^{N} \frac{IIF_i}{N}$, is used as criterion for choosing the most appropriate STF. The logic underling $\overline{TIF}$ is that the higher the number of items included in the test, the higher the penalization of the TIF. In other words, for each θ level and given the same IIFs, STFs composed of fewer items would be favored.

Item Selection Procedures

This section presents two item selection procedures: (i) the *item locating algorithm* (ILA), which is the procedure aimed at recovering a specific TIF determined a priori by locating the most appropriate items from the item bank, and (ii) the *brute force procedure* (BFP), which selects the STF that best recovers the TIF determined a priori after trying all the possible combinations of items of different lengths.

Both ILA and BFP require the definition of a TIF target, which describes the desired characteristics of the STF, regardless of its length. Both procedures attempt at recovering as best as they can the TIF target, denoted as TIF^*. The BFP was introduced to check whether ILA is

able to recover the best possible combination of items among all the possible combinations of items of different lengths that can be obtained from the item bank. The simulation study compares the performance of ILA and that of BFP in 100 simulations. Further details are provided in Section "Simulation Study".

Item Locating Algorithm

ILA is an iterative algorithm that compares the TIF^* with temporary TIFs, denoted as TIF_{TE}, obtained at each iteration by adding one item at the time. ILA iterates until the absolute difference between TIF^* and the TIF_{TE} that includes the last selected item is equal to or greater than the difference between the TIF^* and the TIF_{TE} that does not include the last selected item (i.e., termination criterion). Notably, TIF_{TE} is the mean temporary TIF obtained as $TIF_{TE} = \frac{\sum_{i \in Q} IIF_i}{||Q||}$, where $||Q||$ denotes the cardinality of the vector Q that contains the indexes of the items selected for inclusion in the STF up to that iteration.

The algorithm iterates the following steps until the termination criterion is reached:

1. The absolute difference between TIF^* and TIF_{TE} is computed, $\Delta_{TIF} = |TIF^* - TIF_{\text{TE}}|$. Since at the beginning no items are selected for inclusion in the STF, $Q = \emptyset$ and $||Q|| = 0$, so $\Delta_{TIF} = |TIF^* - 0|$.

2. A θ_{target} is determined as the θ level for which the absolute distance between the TIF^* and TIF_{TE} is maximum, $\arg\max \Delta_{TIF}$.

3. The index of the first item to be included in Q is the index of the item whose location is closest to the θ_{target}, $\text{argmin}_{i \in \{1,\ldots,N\} \setminus Q} |\theta_{target} - b_i|$.

4. The mean temporary TIF is computed as $TIF_{TE} = \frac{\sum_{i \in Q} IIF_i}{||Q||}$

5. Repeat from Step 1 until the termination criterion is reached.

The Brute Force Procedure

BFP develops the STFs considering all the possible combinations of items without repetition and compares the STFs from those of length 1 (i.e., those composed of one item only) to those of length L, where $L = N - 1$ and N is the number of items included in the item bank. For each of the possible lengths $l = \{1, \ldots L\}$ of the STF, a $\binom{N}{l}$ number of STFs is developed. For instance, if the item bank is composed of $N = 10$ items, $l = \{1, \ldots, 9\}$, and the BFP will develop $\binom{10}{1} + \binom{10}{2} + \binom{10}{3} + \binom{10}{4} + \binom{10}{5} + \binom{10}{6} + \binom{10}{7} + \binom{10}{8} + \binom{10}{9} = 1,022$ STFs.

For each of the possible combinations of items of length l, the $\overline{TIF}$ is computed, and the absolute difference from the TIF target, $\Delta_{TIF} = TIF^* - \overline{TIF}$, and a mean of the absolute difference are obtained $\overline{\Delta}_{TIF}$. The best STF is the one with the lowest value of $\overline{\Delta}_{TIF}$, that is the one that presents the lowest absolute distance from the TIF target.

Simulation Study

The simulation study was run in R (R Core Team, 2022), and it iterated the following steps for 100 times:

1. Generate an item bank composed of $N = 6$ items with difficulty parameters drawn from a uniform distribution $\mathcal{U}(-3, 3)$ and discrimination parameters drawn from a uniform distribution $\mathcal{U}(.90, 2.0)$;

2. Generate a mean TIF target by randomly selecting items from the item bank. The number of selected items varied between 2 and 5 (Mean $= 3.34 \pm 1.13$). The starting difficulty and discrimination parameters of the selected items are modified by adding or subtracting values randomly extracted, at each iteration, from uniform distributions $\mathcal{U}(-0.20, 0.20)$. A constraint is made on the discrimination parameters such that they cannot result in negative values after they are modified through the addition or subtraction of the value;

3. Considering the TIF defined at Step 2 and the item parameters generated at Step 1:

    - ILA searches for the best possible STF able to recover the TIF target
    - BFP looks for the best possible item combination. Given that the item bank is composed of $N = 6$ items, $L = 5$ and $\binom{6}{1} + \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} = 62$ STFs are developed and compared.

To evaluate the performance of ILA, the following criteria have been followed:

- Symmetric distance between the indexes of the items selected by BFP and ILA. On the ground of the symmetric distance, it is possible to compute the sensitivity (i.e., the number of items that are identified by BFP and selected by ILA), specificity (i.e., the number of items that have been excluded by both BFP and ILA), and accuracy (the sum between specificity and sensitivity divided by the length of the item bank).

- Rank of the STF obtained with ILA in the distribution of STFs obtained with BFP, ordered according to their increasing distance from the average TIF of the STFs that best resemble the TIF target.

Moreover, the occurrences in which the cardinalities of Q were equal according to both ILA and BFP $||Q_{\text{ILA}}|| = ||Q_{\text{BFP}}||$ were counted in each iteration. If $||Q_{\text{ILA}}|| = ||Q_{\text{BFP}}||$, then we checked whether the two strategies resulted in the same selection of items. Otherwise, the difference between the cardinalities was computed as $||Q_{\text{ILA}}|| - ||Q_{\text{BFP}}||$ and it was averaged across the iterations.

Results

Among the 100 simulations, ILA and BFP resulted in STFs of same lengths in 57 cases, out of which 41 cases resulted in the same selection of items. In the 34% of cases, ILA selected less items than BFP, while it selected more items than BFP in the 9% of cases. The weighted mean of item difference between ILA and BFP was 0.77.

Averaging across the 100 iterations, ILA included the same items as BFP (sensitivity) and excluded the same items as BFP (specificity) in the 72% and 85% of the occurrences, respectively.

Concerning the percentile rank of the distance from the best STFs identified by BFP, in the 74% of instances ILA produced STFs within the 10nth percentile of distance, while the median percentile rank was 3.17. This means that in the 50% of the cases, ILA developed a STF that was within the fourth percentile of distance from the best possible STF.

Final Remarks

As suggested by the results of the simulation study, ILA appears to be a promising algorithm for shortening tests in an IRT framework. Nonetheless, work is still needed. As it is now, ILA grounds item selection only on the item location with respect to the identified θ target. Future developments of the algorithm should consider the discrimination parameter as well, possibly by grounding the item selection on their information functions with respect to the θ target.

References

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

Epifania, O. M., Anselmi, P., & Robusto, E. (2022). Pauci sed boni: An item response theory approach for shortening tests. In *The annual meeting of the psychometric society* (pp. 75–83). doi: https://doi.org/10.1007/978-3-031-27781-8_7

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`