Nothing lasts forever – only item administration: Item Response Theory procedures to shorten tests

Ottavia M. Epifania[1] and Livio Finos[2]

[1] Department of Psychology and Cognitive Science, University of Trento, IT
ottavia.epifania@unitn.it
[2] Department of Statistics, University of Padova, IT

Abstract. da rivedere Although a larger number of test items improves measurement validity, the effect of respondents' fatigue on the response quality should be acknowledged for developing reliable measurement tools. This contribution presents an item response theory-based algorithm (denoted as Léon) able to shorten existing tests by concurrently accounting for the measurement precision of the abbreviated test and the tiredness of the respondents, which is here conceptualized as the probability of observing careless errors as the number of administered items increases. A simulation study compares the performance of Léon of approximating the measurement precision that would be obtained from the full-length test without the effect of the tiredness against that of another algorithm that does not account for the tiredness of the respondents. Although on average the two algorithms select the same number of items, Léon provides a better approximation to the measurement precision of the full-length test than the other algorithm.

Keywords: Item response theory, careless error, information functions, short test forms

1 Introduction

As a general rule of thumb, the higher the number of items in a test, the better the measurement in terms of validity and reliability. However, there is a trade-off between the number of administered items and the response quality. As such, the trade-off between the number of administered items and the tiredness of the respondents should be kept in mind to obtain reliable and precise measurement tools. Item Response Theory (IRT, see, e.g., [1]) provides an ideal framework for shortening existing tests (or for developing tests from item banks) given the detailed information that they provide of the measurement precision of each item with respect to different levels of the latent trait. In this contribution, we present a new IRT-based algorithm for developing short test forms (STFs) from existing tests, denoted as Léon. This algorithm accounts for the tiredness of the respondents during the item inclusion process, such that it attempts at minimizing the number of selected items while accounting for the tiredness of the

respondents in order to maximize the measurement precision (as expressed by the test information function, TIF) of the STF. The tiredness of the respondents has been here conceptualized as the careless error related to the rank of the items during the administration.

2 Item Response Theory and Information Functions

In IRT models for dichotomous responses (e.g., correct vs. incorrect), the probability of observing a correct response on item i by person p depends on both the charcteristics of the respondent (as described by their latent trait level, θ_p) and on the characteristics of the item, which can be described by different parameters. IRT models differentiate according to the number of parameters used for describing the characteristics of the items. According to the 4-Parameter logistic model (4-PL, [2]), the probability of a correct response can be formalized as:

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

where θ_p is the latent trait level of person p, b_i is the location of the item on the latent trait (i.e., difficulty parameter, the higher the value, the higher the difficulty of the item), and a_i describes the ability of i to discriminate between respondents with different latent trait levels (i.e., discrimination parameter, the higher the value, the higher the discrimination ability of the item). Parameters c_i and d_i represent the lower and upper asymptotes of the probability, which should naturally tend to 0 and 1 when $\theta \to -\infty$ and $\theta \to +\infty$, respectively. The pseudoguessing parameter c_i describes the probability of observing a correct response on i for $\theta \to -\infty$, such that the probability tends to c_i instead of 0. The d_i parameter describes the probability of observing a correct response on i for $\theta \to +\infty$, such that this probability tends to d_i instead of 1. The careless error probability is $1 - d_i$ (i.e., the probability of not endorsing the item given that the $\theta_p > b_i$).

In this study, we operationalized the tiredness of the individuals as the careless error probability $1 - d_i$, which increases as the administration goes on. As $1-d_i$ increases, the upper aymptote d_i decreases. This probability is instrinsically related to the rank of the item in the test, hence the careless error parameter in Equation (1), which is usually a property of the item (d_i), becomes a property of the item rank $r = \{1, \ldots, R\}$ in the test, d_r, with the constraint $d_{r-1} > d_r$.

2.1 Information Functions

The measurement precision of each item with respect to different levels of the latent trait can be expressed by the *item information functions* (IIFs). Given that in this application the tiredness of the respondents has been conceptualized as the careless errors observed as the administration goes on, the typical formalization of the IIF for the 4-PL [3] has been modified to account for the rank of presentation r of item i, as follows:

$$\mathrm{IIF}_i(r) = \frac{a^2[P(\theta) - c_i]^2[d_r - P(\theta)]^2}{(d_r - c_i)^2 P(\theta)Q(\theta)}, \tag{2}$$

whereas in the typical formalization $d_r = d_i$. The informativeness of each item is strongly influenced by the location of the item on the latent trait with respect to a specific latent trait level θ, its discriminativity, and the probability of lucky guess and careless error. In absence of lucky guess and careless error, the IIF is maximum when $\theta_p = b_i$, and it decreases as the distance between b_i and θ_p increases. Moreover, the higher the discrimination of the item a_i, the more informative the item. When the lucky guess and careless error are taken into account, the informativeness of the item decreases, the IIF is lower, and it maximum is shifted.

The *test information function* (TIF) is the sum of the IIFs, such that its shape (i.e., its informativeness with respect to different latent trait levels) and its height (i.e., the amount of information for different latent trait levels) depend on the items distribution along the latent trait. The more the items with high discrimination and low lucky guess and careless error parameters are spread throughout the latent trait, the more the test is informative of different regions.

3 Item Selection Algorithms

The two algorithms aim at reducing the distance between a target TIF (TIF-target, describes the desired measurement precision of the STFs) and a provisional TIF (pTIF) obtained from the items included in the STF. At each iteration, an item is included in the STF according to its ability of bridging the gap between the two TIFs. The algorithms stop when the distance between the TIF-target and the pTIF with the last considered item is equal to or greater than the distance between the TIF-target and the pTIF without the last considered items (i.e., termination criterion).

In what follows, the TIF-target is represented by the TIF that would be obtained if respondents would never get tired during the administration of all the items in a test B. Therefore, the TIF-target is denoted as TIF_B.

Given that B is the set of item without the effect of the tiredness, $d_i = 1$, $\forall i \in \{1, \ldots, I\}$. To include the tiredness of the respondents, the vector of careless error parameters d_r should be modified considering the item rank of presentation. In this application, d_r is modified with an exponential function, $d_r = \exp(-\lambda r)$ (where λ is the non-negative speed parameter that determines the steepness of the function and r is the rank of the r-th item in the administration). of the i-th item in the administration). The set of items with d' i is denoted B'.

Since the TIF increases as the number of items in a test increases, the comparison between TIF_B and the TIF of the STFs is based on the mean TIF (i.e., the TIF divided by the number of items). Nonetheless, in what follows the mean TIF will be simply referred to as TIF_{Q_x}, with $x \in \{B, B', \mathrm{Frank}, \mathrm{Léon}\}$ and Q is the set of items in the tests B or B' or the subset of items included in the STFs generated by Frank and Léon.

3.1 Frank

Frank considers the entire latent trait for the item selection, in that it selects the item whose IIF is best able to reduce the distance from the TIF_B along the entire latent trait, as follows:

At $k = 0$: IIF_i, $\text{TIF}^0(\theta) = 0\,\forall\theta$, $Q^0 = \emptyset$. For $k \geq 0$,

1. $A^k = B \setminus Q^k$
2. $\forall i \in A^k$, $p\text{TIF}_i^k = \frac{\text{TIF}^k + \text{IIF}_i}{||Q^k||+1}$, with $c_i = 0$ and $d_i = 1$, $\forall i \in B$
3. $i^* = \arg\min_{i \in A^k} |\text{TIF}_B - \text{pTIF}_i^k|$
4. Termination criterion: $|\text{TIF}_B - \text{pTIF}_{i^*}^k| \geq |\text{TIF}_B - \text{TIF}^k|$:
   - FALSE: $Q^{k+1} = Q^k \cup \{i^*\}$, $\text{TIF}^{k+1} = p\text{TIF}_{i^*}$, iterates 1-4
   - TRUE: Stop, $Q_{\text{Frank}} = Q^k$

At $k = 0$, the subset of items Q^0 is empty and the TIF^0 is 0 for all the θ levels. At each iteration k: (1.) a set of available items is generated as the items in the item bank that have not been included in the STF yet, $A^k = B \setminus Q^k$; (2.) An average provisional TIF, pTIF, is computed by adding the IIF of each of the items in the set of the available items A^k, one at the time, to the TIF obtained from the items in Q^k (The denominator is obtained by adding 1 to the cardinality of Q^k); (3.) Among all the items in A^k, the one that allows for minimizing the distance between pTIF and TIF_B is included in i^*; (4.) The termination criterion is tested. If the distance between the TIF_B and the pTIF_{i^*} is greater than or equal to the distance between the TIF_B and the TIF^k (i.e., the TIF obtained from the items in the subset Q^k, without item i^*) (TRUE), then the item i^* does not contribute in the reduction of the distance from the TIF_B, the algorithm stops, and the final item selection is the one without the item in i^*, $Q_{frank} = Q^k$. Conversely (FALSE), the item in i^* does contribute in the reduction of the distance from the TIF_B, hence it is included in the set of items and a new iteration starts, $Q^{k+1} = Q^k \cup \{i^*\}$.

3.2 Léon

Differently from Frank, at step (2.) Léon computes the IIF_i by considering the d_r related to the number of items included in the STF up to that point.

4 Simulation study

4.1 Simulation design

The procedure is replicated 100 times. At each replication, a test B of 50 items with difficulty ($b_i \sim \mathcal{U}(-3, 3)$) discrimination ($a_i \sim \mathcal{U}(.90, 2)$) parameters drawn from uniform distributions is generated. Lucky guess and careless error parameters are constant, $c_i = 0$ and $d_i = 1$, $\forall i \in B$. The TIF_B (i.e., the TIF-target) is obtained as the average TIF from the items in B, and describes the measurement

precision that would be obtained if respondents were administered with all the items without getting tired.

At each replication, a new test, B', is generated, where $b'_i = b_i$, $a'_i = a_i$, $c_i = 0$, $\forall i \in B'$, and $d_r = \exp(-\lambda r_i)$, with $\lambda = 0.01$ and $r = \{0, \dots, ||B|| - 1\}$.

At each replication, Frank and Léon generate a STF for approximating the TIF_B. An average TIF is computed considering the item in B', denoted as $\text{TIF}_{B'}$. Although Frank grounds the item selection on the set B, the final TIF is computed by considering the vector d_r associated to the rank of the items in Q_{Frank}.

4.2 Comparison

The average distance from TIF_B of the STFs generated by Frank and Léon at each iteration, as well as of the test obtained from B' has been considered as a criterion for evaluating the ability of Léon of generating informative STFs while accounting for the tiredness of the respondents, $\Delta_x = |\text{TIF}_B - \text{TIF}_{Q_x}|$, with $x \in \{B', \text{Frank}, \text{Léon}\}$. Trivially, when the $\text{TIF}_{B'}$ is considered, $Q = B'$.

To better understand whether acknowledging the tiredness of the respondents influences the number of items included in Q, the cardinalities of the STFs generated by Frank and Léon, $||Q_{\text{Frank}}||$ and $||Q_{\text{Léon}}||$, respectively, have been compared.

5 Results

On average, Léon and Frank included the same number of items in the STF, $M_{\text{Léon}} = 6.75 \pm 3.06$, $M_{\text{Frank}} = 6.31 \pm 2.40$, $t = 1.13$, $df = 187.59$, $p = .26$, suggesting that accounting for the tiredness of the respondents does not influence the number of items included in the STF.

Figure 1 illustrates the distributions of the distances from TIF_B (y-axis) of Q_x (x-axis).

Overall, $\text{TIF}_{B'}$s are the most distant from TIF_B, while $\text{TIF}_{\text{Léon}}$s are the closest ones. Frank falls in between the two, also presenting the least consistent performance across the replications. dovrei aggiunger some sort of test

6 Final Remarks

This manuscript presented an IRT-based algorithm, denoted as Léon, for the generation of STFs. Léon grounds the item selection by concurrently considering the number of items included in the STF and the measurement precision with respect to an ideal target where all the items are administered in absence of careless error (i.e., TIF-target). The ability of Léon of approximating the TIF-target has been compared against that of another algorithm (Frank) not accounting for the tiredness of the respondents during item selection and with the measurement precision obtained from the administration of the entire test, including the careless mistakes due to the response fatigue.

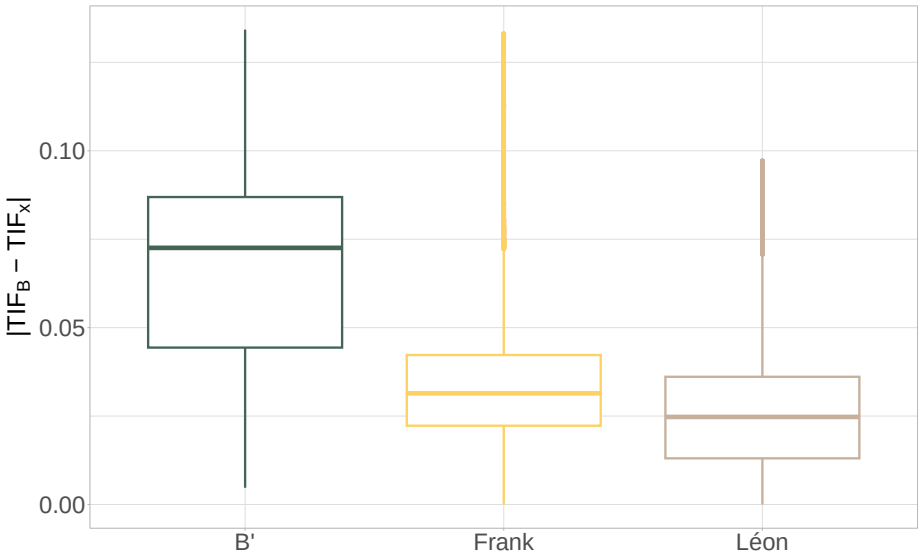

Fig. 1. Distance from TIF_B of $\text{TIF}_{B'}$, $\text{TIF}_{\text{Frank}}$, and $\text{TIF}_{\text{Léon}}$ across the 100 replications.

Generally, the results suggest that the administration of fewer items provide better measurement tools than administering the entire test, especially if the selection process accounts for the effect of the response fatigue.

Although the results are promising, several limitations must be acknowledged. Firstly, while this study primarily focuses on approximating ideal measurement precision as expressed by the TIF-target, no investigations on the precision with which the latent trait is estimated with the STF are presented. Future studies should further investigate this issue to understand whether administering less items brings actual benefits in the estimation of the latent trait. Secondly, the operalization of the tiredness of the respondents as an increase of the probability of committing careless error as the administration goes on is an arbitrary choice, and other possibilities should be investigated.

References

1. Baker, F. B., Kim, S.-H. (2017). The basics of Item Response Theory using R. Springer.
2. Barton, M. A., Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. Princeton, NJ: Educational Testing Service.
3. Magis, D. (2013). A note on the item information function of the four-parameter logistic model. Applied Psychological Measurement, 37(4), 304-315.