How much is too much? 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

a me questa introduzione piace, mi dispiace toglierla 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, baker2017) 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 ability of Léon of developing STFs able to approximate the TIF that would be obtained by administering all the items under the assumption that respondents would never get tired is investigated in a simulation study. Specifically, Léon's ability is compared against that of another algorithm for developing STFs, which does not account for the tiredness of the respondents. 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), 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), 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), and c_i and d_i describe the probability of observing a correct response when $\theta \to -\infty$ and $\theta \to +\infty$, respectively. When $\theta \to -\infty$, the probability of observing a correct response should tend to 0. Likewise, when $\theta \to +\infty$, the probability of observing a correct response should tend to 1. However, there might be instances where respondents with θ levels below the difficulty of the item, whom are hence expected not to respond correctly, might provide the correct response out of luck. The pseudoguessing parameter c_i describes this probability, such that the probability of observing a correct response on i for $\theta \to -\infty$ tends to c_i instead of 0. The same but inverse consideration applies when $\theta \to +\infty$. The d_i parameter describes the probability of endorsing the item given that the latent trait is above the location of the item, such that the probability of observing a correct response on i for $\theta \to +\infty$ tends to d_i instead of 1.

By constraining $\forall i \in B$, $d_i = 1$ (where B is the set of items in a test) in equation 1, the 3-Parameter logistic (3-PL) model is obtained. From the 3-PLmodel, the 2-Parameter logistic model (2-PL) is obtained by constraining $\forall i \in B$, $c_i = 0$, and the 1-parameter logistic model (1-PL, equivalent to the Rasch model) is obtained by constraining $\forall i \in B$, $a_i = 1$.

In this study, we operationalized the tiredeness of the individuals as the probability of committing careless error as the administration goes on, hence as a function of the number of items in a test and their relative position. This probability is instrinsically related to the rank of the item in the test, and it increases exponentially as a function of the number of administered items. As such, the the careless error parameter in Equation 1, which usually a property of the item per se expressed as d_i, in this case becomes a property of the item rank $r = \{1, \ldots, R\}$ in the test, d_r, with the constraint that $d_{r-1} < d_r$.

2.1 Information Functions

The meausemrement precision of each item with respect to different levels of the latent trait can be expressed by the so-called *item information functions* (IIFs), which for the 4-PL is typically formalized as:

$$IIF_i = \frac{a^2[P(\theta) - c_i]^2[d_i - P(\theta)]^2}{(d_i - c_i)^2 P(\theta)Q(\theta)}. \tag{2}$$

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 reaches its maximum when the location of the item b_i matches the latent trait level θ, and it decreases as the distance between b_i and θ increases. Moreover, the higher the discrimination of the item a_i, the more informative the item. However, when the lucky guess and careless error are taken into account, the informativeness of the item decreases, and, most importantly, the maximum of the IIF does not corresponds to the item location on the latent trait and it is generally lower.

The *test information function* (TIF) is the sum of the IIFs, such that its shape (i.e., its informativeness with respect to different levels of the latent trait) and its height (i.e., the amount of information for different levels of the latent trait) depend on the distribution of the items 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 would be informative of different regions. The more the items have lucky guesses and careless errors, the less the TIF.

Again, given that in this application the tiredness of the respondents has been conceptualized as the probability of observing careless errors as the administration goes on, the d parameter in Equation 2 is a property related to its order of presentation in the test r, and it is hence denoted as d_r with $r = \{1, \ldots, R\}$.

3 Item Selection Algorithms

The two algorithms are based on the same principle of reducing as much as possible the distance between a TIF-target and a provisional TIF (pTIF) obtained from the items selected up to that point. At each iteration, the item is included

in the STF according to its ability of bridging the gap between the two TIFs. The algorithms stop when the last item considered for inclusion in the STF does not contribute to bring the pTIF closer to the TIF-target, that is 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 obtained on all the items in a test without careless error parameters, that is the TIF that would be obtained if respondents would never get tired during the administration. Given that the original set of items is denoted as B, the TIF-target obtained from this set is denoted as TIF_B.

The parameters of the items in B are defined through an $I \times 4$ matrix, where I is the total number of items in B ($||B||$, where $||X||$ denotes the cardinality of set X) and the 4 columns contain the item parameters b_i, a_i, c_i, and d_i. Given that B is the set of item without the effect of the tiredness (i.e., without careless parameters), $d_i = 1$, $\forall i \in \{1, \ldots, I\}$. To include the tiredness of the respondents, the vector of careless error parameters d_i is modified with an exponential function, $d'_i = \exp(-\lambda r_i)$ (where λ is the speed parameter that determines the steepness of the function and r is the rank 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 STF obtained by the item selections provided by the two algorithms is based on the mean TIF (i.e., the TIF divided by the number of items in the STF). Nonetheless, in what follows the mean TIF will be simply referred to as TIF_x, with $x \in \{B, B', \text{Frank}, \text{Léon}\}$.

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:

Non è propriamente così, frank calcola tutte le IIF all'inziio considerando i parametri degli item in B' e se le porta avanti ma quindi dovrei creare un altro oggetto che contiene tutte le iif i? ci provo tanto ormai è tutto una merda

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

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

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, $Q^{k+1} = Q^k \cup \{i^*\}$.

4 Simulation study

4.1 Simulation design

The procedure is replicated 100 times. At each replica, a test of 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, $c_i = 0, \forall i \in B$ and $d_i = \forall i \in B$. The TIF_B 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.

A new test B' is generated by keeping the item parameters equal to those in B, with the only exception of the careless error parameters, which are included as related to the rank of the item in the item bank with an exponential function. The assumption is that the first administered item does not have a careless error probability ($d_{1^{st}} = 1$), but it increases as the administration goes on. As such, the careless error probabilities associated to the second and third administered items are, respectively, $d_{2^{nd}} = .99$ and $d_{3^{rd}}.98$. In this application, the maximum careless error probability associated to the last administered item is $d_{50^{th}} = .61$.

At each replication, Frank and Léon generate a STF for approximating TIF_B.

4.2 Comparison

The performance of Léon has been compared The ability of Frank and Léon in reducing the distance from the TIF_B has been considered. Moreover, the TIF of the items in B' has been computed as well and compared against those of the STFs resulting from the application of Frank and Léon. The rationale is as follows. The more the item in a test, the higher the information for different regions of the latent trait. However, administering too many items might be an highly demanding task, such that the respondent might get tired and their

response accuracy might decrease during the administration, such that the last administered items might include error variance not related to the construct under investigation. In this light, it should be more convenient to administer less but highly informative items, able to approximate a specific TIF target, than to administer the entire test to obtain precise measurements of the latent trait.

5 Results

On average, Léon and Frank included the same of items included more items in the STF than Frank did, $M_{\text{Léon}} = 6.75 \pm 3.06$, $M_{\text{Frank}} = 6.31 \pm 2.40$, $t = 1.13$, $df = 187.59$, $p = .26$, 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 of $\text{TIF}_{B'}$ (green dots), TIF_{Frank} (yellow triangles), and $\text{TIF}_{\text{Léon}}$ (brown squares).

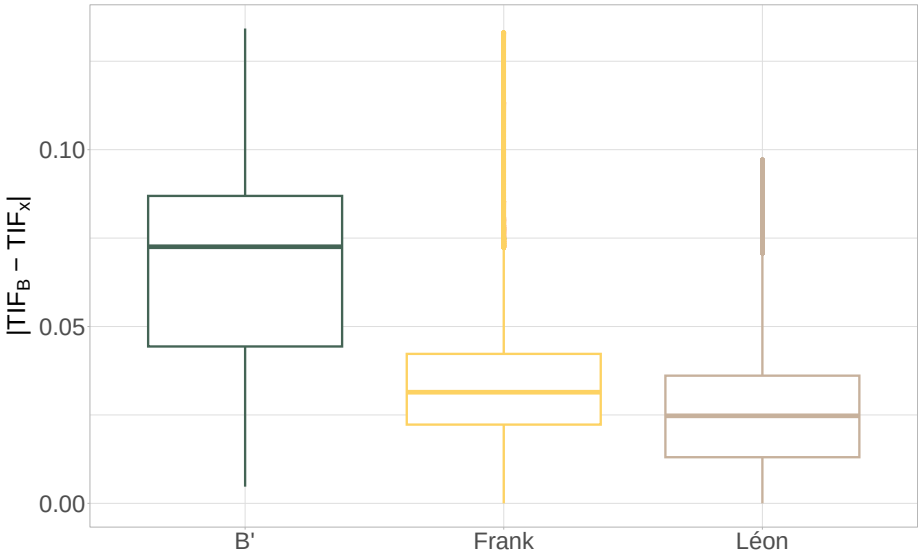

Fig. 1. Distance from TIF_B of $\text{TIF}_{B'}$ (green dots), TIF_{Frank} (yellow triangles), and $\text{TIF}_{\text{Léon}}$ (brown squares) in each of the 100 replications. The horizontal green, yellow, and brown lines are the average distance from TIF_B of $\text{TIF}_{B'}$, TIF_{Frank}, and $\text{TIF}_{\text{Léon}}$, respectively

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.

6 Final Remarks

This manuscript presented a first attempt at the development of an IRT-based algorithm, denoted as Léon, for the generation of informative and static STFs able to account for the tiredness of the respondents. In the item selection for inclusion in the STF, Léon considers the number of items included up to that iteration and adds a penalty in terms of higher probability of careless error, which increases as the number of items included in the STF increases. Its performance in approximating a TIF target, which is here conceptualized as the TIF that one would obtain considering the entire administration of the test if respondents would never get tired (i.e., without careless error) , is compared against that of another IRT-based algorithm, denoted as Frank. Differently from Léon, Frank does not add any penalization for the number of items included in the STF. Finally, the distance between the TIF obtained from the administration of the entire test with and without tiredness has been considered as well.

The results of a simulation study suggest that the administration of fewer items selected also considering the tiredness of the respondents might provide better measurement tools than administering the entire test and tiring out the respondents.

Although the results are promising, there are several limitations that should be acknowledged. Firstly, this study is focused on the approximation of the TIF target. However, the final aim with which tests are administered is to estimate the latent trait of the respondents. The lack of the precision of estimation of the latent trait of the respondents represents the main limitation and future studies should focus on this issue. 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.

References

1. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. 147, 195?197 (1981). doi:10.1016/0022-2836(81)90087-5
2. May, P., Ehrlich, H.-C., Steinke, T.: ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148?1158. Springer, Heidelberg (2006). doi:10.1007/11823285_121
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid information services for distributed resource sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181?184. IEEE Press, New York (2001). doi:10.1109/HPDC.2001.945188
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The physiology of the grid: an open grid services architecture for distributed systems integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov