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] Université de Paris-Sud, Laboratoire d'Analyse Numérique, Bâtiment 425, F-91405 Orsay Cedex, France

Abstract. The abstract should summarize the contents of the paper using at least 70 and at most 150 words. It will be set in 9-point font size and be inset 1.0 cm from the right and left margins. There will be two blank lines before and after the Abstract. ...

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 (giuro che ho della letteratura in merito). Quindi non torturiamo le persone che dopo di un po' non ne possono più. Item Response Theory (IRT) provides an ideal framework for shortening existing tests (or for developing tests from item banks) given the detailed information that they provide with respect to the measurement precision of each item considering different levels of the latent trait. In this contribution, we present a new algorithm for shortening tests that take into accounts the number of administered items by considering the "tiredness" of the respondents for each of the items included in the short test form (STF). The performance of the algorithm in retrieving an ideal target information function (i.e., the measurement precision that would be obtained by administering all the items in a test if the respondents would never get tired) is investigated in a simulation study. il concetto c'è mancano le parole

2 Item response theory e mortacci

In Item Response Theory (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 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 lucky guess parameter c_i describes the probability of endorsing the item even if the θ level is below the difficulty of the item, such that the probability of observing a correct response for $\theta \to -\infty$ tends to c_i instead of 0. The same but inverse consideration applies for the upper asymptote that describes the probability of giving the correct response for $\theta \to +\infty$. The d_i parameter describes the probability of not endorsing the item given that the latent trait is above the location of the item, such that the probability of observing a correct response 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 set), the 3-Parameter logistic (3-PL) model is obtained. From 3-PL, 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 $\forall i \in B$, $a_i = 1$.

2.1 Information Functions

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

Formula del 4-PL

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 discirminativity, 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 on 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.

3 Simulation study

In this study, we operationalized the tiredeness of the individuals as the probability of committing careless error ($1 - d_i$ QUESTA è LA PROBABILITà DI NON CARELESS ERROR DEVO SPIEGARE MEGLIO, mi sto incansiandno basta che alla fine sia coerente), such that the more the items in a test, the higher the probability of careless error. The careless error probability is instrinsically related to the rank of the item in the test, and it increases as the position of the item in the test increases. It increases exponentially with the function SCRIVERE LA FUNZIONE, such that the first administered item has $d_i = 1$ (i.e., does not have a probability of observing an incorrect response given the latent trait level θ).

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 $d_i = 0, \forall i \in B$. The TIF^* is generated as the mean TIF of all items in B, which describes the ideal TIF that one would obtain if respondents are 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 the function SCRIVERE LA FUNZIONE. The assumption is that the first administered item does not have a careless error probability ($d_{1\text{st}} = 1$) and it increases as the administration goes on. As such, the careless error probabilities associated to the second and third administered items are, respectively, .99 and .98. In this application, the maximum careless error probability associated to the last administered item is .61.

Frank and Léon are applied to generate a STF able to recreate the TIF^*. Both algorithm are applied considering the items in B'. While both algorithms are able to consider the careless error parameters in the computation of the IIFs,

only Léon actually accounts for the number of items included in the STF. NON SO COSA STO DICENDO. devo controllare il codice perché ora non capisco. se anche frank calcola la iif considerando la careless error non capisco in che modo dovrebbe differenziarsi da leon ma effettivmente i risultati sono peggiori.

Va detto da qualche parte che:

- La stanchezza dei soggetti viene operazionalizzata come probabilità di non careless error (va beh l'asintoto)

- la tif target è definita come forma ideale dell'informatività che si avrebbe se le persone non si stancassero mai

- la funzione di stanchezza è quella che mi ha dato chatgpt

- va descritto l'algoritmo e va descritto anche Frank (Frank ha solo un passaggio in meno rispetto a Leon perché non considera la stanchezza, basta descrivere leon e poi dire che Frank fa la stessa roba ma non mette la penalizzazione)

4 Results

Table 1. questa è una tabella, cominciamo

Procedure	M	Min	Max
All items	0.07 ± 0.03	< 0.005	0.13
Frank	0.03 ± 0.03	< 0.001	0.14
Léon	0.03 ± 0.02	< 0.001	0.1

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